DOCUMENT RESUME

ED 193 286 TM 800 601

AUTHOR Caulley, Darrel N.
TITLE The Quantitative and Qualitative in the Physical Sciences and the Implications for Evaluation. Research on Evaluation Program, Paper and Report Series, No. 25.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO REL-25
PUB DATE Sep 79
NOTE 119p.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Data: *Measurement Techniques: Number Concepts: *Physical Sciences: Program Evaluation: *Social Sciences
IDENTIFIERS Counting: *Qualitative Data: *Quantitative Data: Ranking

ABSTRACT

Significant questions are addressed in an extensive discussion of the differences between qualitative and quantitative concepts and measurement strategies in the physical sciences. Also included is a discussion of number-generating activities often grouped in the social sciences under the term of measurement. Implications for the redirection of evaluation practice are considered. Specifically, Part I of the report distinguishes between the different types of concepts and the data associated with them. One conclusion is that the initial understanding of a phenomenon must be through qualitative concepts, and from them quantitative concepts may evolve. Part II examines various ways in which numbers are assigned: concluding that neither assignment nor measurement is synonymous with quantification. Part III examines the history of both the qualitative and quantitative in the physical sciences, and the implications for evaluation. The main idea of Part III is that much qualitative work has been prerequisite to fruitful quantification in the physical sciences. Because evaluation draws on the social sciences, which are in early developmental stages, quantification in evaluation may not be as fruitful as qualitative methodology. (Author/GK)

paper and report series

[redacted]

Research on Evaluation Program

Northwest Regional Educational Laboratory
710 S.W. Second Avenue
Portland, Oregon 97204
Telephone: (503) 248-6800

INTERIM DRAFT

Do not cite or quote without author's permission.

Author welcomes reactions and suggestions.

No. 25 THE QUANTITATIVE AND QUALITATIVE IN THE PHYSICAL SCIENCES AND THE IMPLICATIONS FOR EVALUATION

DARREL N. CAULLEY

Northwest Regional Educational Laboratory

September 1979

Nick L. Smith, Director
Research on Evaluation Program
Northwest Regional Educational Laboratory
710 S.W. Second Avenue, Portland, Oregon 97204

Published by the Northwest Regional Educational Laboratory, a private nonprofit corporation. The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

PREFACE

The Research on Evaluation Program is a Northwest Regional Educational Laboratory project of research, development, testing, and training designed to create new evaluatiom methodologies for use in education. This document is one of a series of papers and reports produced by program staff, visiting scholars, adjunct scholars, and project collaborators--all members of a cooperative network of colleagues working on the development of new methodologies.

What is the distinction between qualitative and quantitative approaches in the physical sciences? How are these approaches used in evaluation, viewed here as a subsystem of the social sciences. This report addresses these and other significant questions in an extensive discussion of the differences between qualitative and quantitative concepts and measurement strategies in the physical sciences. Also included here is a discussion of various number generating activities often grouped in the social sciences under the label of "measurement." Implications for the redirection of evaluation practice are also considered.

Nick L. Smith, Editor
Paper and Report Series

CONTENTS

Page

INTRODUCTION vii

PART I. THE DIFFERENCE BETWEEN THE QUALITATIVE AND THE QUANTITATIVE. 1

Types of Concepts - An Introduction. 1
Classificatory Concepts. 4
Changing a Classificatory Concept into a Comparative Concept. 8
Comparative Concepts 14
Interval Concepts. 17
Ratio Concepts 19
Absolute Concepts. 19

PART II. THE WAYS OF ASSIGNING NUMBERS. 21

Measurement. 22
Introduction to Measurement Via Sets. 22
Fundamental Measurement 24
Basic Procedures for Determining the Empirical Relationships Between Quantities and a Property. 25
The Representation Problem. 30
The Various Definitions of Measurement. 33
The Steps that Lead to the Measurement of a Property--An Historical Example. 39
Using Measurements. 44
What data are needed? - Specification. 46
How are data to be used in various contexts? - Generalizability. 46
How accurate are the data? 49
How should the data be expressed? - The language of communication 56
How are the data to be interpreted?. 57
How are the data to be used in decision making 58
Counting 60
Naming by Numbers. 69
Numerical Judgment 71
Indices. 75
Ranking. 88
Number Assignment and Type of Concept. 89
Comparison Between Pbysical and Social Sciences. 90

CONTENTS

Page

PART III. THE HISTORY OF THE QUANTITATIVE AND QUALITATIVE IN THE PHYSICAL SCIENCES AND THE IMPLICATIONS FOR EVALUATION. 92

The View of Quantification from Science Texts. 94
Motives for Normal Measurement 98
The Effects of Normal Measurement. 103
Revolutionary Science. 104
Measurement in the Development of Physical Science 105

INTRODUCTION

Physics, followed by chemistry, is regarded as the most basic physical sciences. There are no clear boundary lines between these two sciences. Physics and astronomy have become highly intertwined and chemistry is important to geology. There has arisen a number of hybrid sciences such as biochemistry and psychophysics. The lines of demarcation of the physical sciences are not clear. For the sake of simplicity, the physical sciences will sometimes be referred to as science. While most examples will be drawn from the physical sciences, an occasional example will be drawn from other sciences.

Within evaluation there has been a certain amount of debate as to the usefulness of quantitative versus qualitative data and the methodology associated with the collection of this data (Parlett and Hamilton, 1972; Stake and Easley, 1978; Guba, 1978). One problem with this debate has been the fact that it is not always clear what the difference is between qualitative and quantitative data; this is the subject of Part I. Data must be about something and one must have a concept about that something in order for the data to be meaningful. Thus the approach of Part I is to distinguish between different types of concepts and hence the data associated with them. One important conclusion is that the initial understanding of a phenomenon must be through qualitative concepts and that, if quantitative concepts evolve, they do so out of these qualitative concepts.

Part II examines the various ways in which numbers are assigned. One conclusion from this part is that number assignment is not synonymous with quantification, since naming by numbers is not a form of quantification. Another conclusion is that measurement is not synonymous

with quantification since there are four ways of assigning numbers which are distinct from measurement, yet are forms of quantification.

Part III examines the history of the quantitative and the qualitative in the physical sciences and looks at the implications for evaluation. A central idea of Part III is that large amounts of qualitative work have usually been prerequisite to fruitful quantification in the physical sciences. Evaluation draws on the social sciences. Since the social sciences are in the early stages of development, this suggests that quantification in evaluation may not be as fruitful as qualitative methodology.

PART I: THE DIFFERENCE BETWEEN THE QUALITATIVE AND THE QUANTITATIVE

One obvious way of distinguishing between the qualitative and the quantitative is to say that the qualitative does not involve the assignment of numbers. Kuhn (1961, p. 32) takes this point of view. However, as will be shown below, numbers may be assigned to the qualitative, but these numbers do not have any essential quantitative meaning. The difference between the qualitative and quantitative can be determined by examining the type of concepts involved.

Types of Concepts--An Introduction

There are five widely used types of concepts that are used in association with the assignment of numbers to the properties of objects and phenomena--classificatory, comparative, interval, ratio, and absolute. While Hempel (1952, p. 50) and Carnap (1962, p. 8) have discussed the first two types, their discussion has been broadened extensively in this discussion. The terms, "interval, ratio, and absolute concepts" are new and were suggested by the notion of interval, ratio, and absolute scales.

To have a classificatory concept is to have a criterion or criteria (which will be a property or properties) which serve for the classification of entities or phenomena into two or more mutually

exclusive kinds. Examples are: male and female; hot and cold; acids, bases, and salts; intelligent and not intelligent.

To have a comparative concept of a property is to envision it as existing in graduations of "more," "less," or "equal" in amounts. For example, a classificatory concept of temperature as hot or cold can be replaced by a comparative concept of temperature by saying that x is warmer than y (or colder, or equally warm, as the case may be).

An interval or ratio concept of a property occurs when it is possible to specify amounts of the property in terms of units. When temperature is specified in degrees centigrade, it is an interval concept. The difference between interval and ratio concepts is that for ratio concepts, there is a true zero point which is not arbitrary. Examples of ratio concepts are length, time, velocity, volume, mass, force, electric charge, and price. Numbers are attached to interval and ratio concepts by means of measurement. Measurement can also be used in relation to undimensional comparative concepts but cannot be used in relation to classificatory concepts, multidimensional comparative concepts, or absolute concepts.

An absolute concept is one that does not involve an arbitrary unit or zero point. Numbers are attached to an absolute concept by enumeration (i.e., counting). When numerosity is conceived as "many" or "few," it is a classificatory concept. When numerosity is conceived as "greater than, less than, or equal in number," it is a comparative concept. When the number of entities in a group are counted, then numerosity is conceived as an absolute concept.

Comparative, interval, ratio, and absolute concepts are known collectively as quantitative concepts, whereas classificatory concepts

are qualitative concepts. Whereas entities, phenomena, and properties* can all be classified, only properties can be quantified. A property that can be quantified is one for which it is possible to say that entity or phenomenon A has more of the property, less of the property, or an equal amount of the property, compared to entity or phenomenon B. A quantitative concept refers to a property that can be quantified.

Concepts can also be classified according to whether they are unidimensional or multidimensional. A multidimensional concept is one that can be broken down into subclassifications or is composed of parts. For example, the concept of intelligence is commonly broken down into verbal intelligence and numerical intelligence. Length, on the other hand, cannot be broken down into subclassifications or parts. It is unidimensional.

Because many of the quantities in the physical sciences are calculated from other quantities, it might appear that they are, therefore, multidimensional. For example, the average density of an object is calculated by dividing the mass of the object by its volume. However, density is not multidimensional in the sense that mass and volume are not subclassifications of density. Density cannot be said to be composed of two parts, mass and volume. In contrast, the concept of the cephalic index used in anthropometry to indicate the shape of the human skull is a multidimensional concept. It is composed of two parts,

*An entity refers to the existence of an object as contrasted with its properties. A phenomenon refers to an object or event known through the senses rather than thought or nonsensuous intuition. A property is a quality or trait belonging and especially peculiar to an object or event.

the length* and breadth* of the human skull. The rule for finding the cephalic index is to divide the shorter of the two measurements by the longer and multiply by 100. The index expresses what percentage the shorter of the two measurements is of the longer measurement with which it is compared. Thus the cephalic index is composed of two parts and is a multidimensional concept.

This discussion has served as an introduction to the various types of concept. What follows is an expanded discussion of them.

Classificatory Concepts

One way developments in knowledge occur is by recognizing objects or phenomena as being essentially the same or different. This leads to the classification of objects or phenomena. A classification of objects or phenomena in a given domain D (such as chemical compounds, psychiatric conditions, animals, plants, stars, etc.) is effected by laying down a set of two or more criteria (attributes, properties) such that each element of D satisfies exactly one of these criteria. Each criterion determines a certain class. If each element in D satisfies exactly one of the criteria, then the classes thus determined are mutually exclusive, and they are jointly exhaustive of D.

The requirements of exclusiveness and exhaustiveness may be satisfied as a logical consequence of the determining criteria or as amatter of empirical fact. An example of the first alternative is the

*The length is measured from the glabella, that is the convex part of the forehead immediately above the root of the nose. The furtherest point from this in the sagittal plane may be the inion, that is the external occipital protuberance, or may lie above this on the interparietal part of the occipital bone. The greatest width is generally that between the two parietal eminences (Stibbe, 1930, p. 179).

classification of human skulls into five classes according to the value of the cephalic index

Class Name	Value of Cephalic Index (c(x))
Dolichocephaly	$c(x) \leq 75$
Subdolichocephaly	$75 < c(x) \leq 77.6$
Mesaticephaly	$77.6 < c(x) \leq 80$
Subbrachycephaly	$80 < c(x) \leq 83$
Brachycephaly	$83 < c(x)$

The requirements of exclusivenss and exhaustiveness are satisfied as a logical consequence of the determining criteria, since any human skull will fall into one class only and every human skull will fall into one of the five classes. This is true also for dichotomous classifications which involve some property and its denial. Examples are the division of integers into those which are and those which are not integral multiples of 2; the division of chemical compounds into organic and inorganic; the division of bacteria into Gram-positive and Gram-negative. The classification of humans into males and females on the basis of primary sex characteristics, the classification of animals on the basis of their morphology, and the classification of crystals on the basis of their structure, are examples which are exclusive and exhaustive empirically and not logically.

The division into classes may be made on the basis of a single criterion or the basis of many. In the above example involving the classification of human skulls, the division into classes was on the basis of a single criterion, the value of the cephalic index. This is irrespective of the fact that the cephalic index is, itself, a multidimensional concept. On the other hand, the classification of

elements into groups in the periodic table by Mendeleev was based on two properties of each element. That is, more than one criterion was used. Group One of the periodic table contains the alkali metals--lithium, sodium, potassium, rubidium, caesium, and francium. No single criterion determines the inclusion of any one of these metals under the category of an alkali metal. The concept of an alkali metal is a multidimensional classificatory concept.

A single property may be conceptualized in more than one of the ways--classificatory, comparative, interval, ratio, and absolute. The classes, hot, warm, and cold involve conceptualizing temperature in a classificatory way. A comparative concept of temperature is involved when it is said that x is warmer than y (or colder, or equally warm, as the case may be). When temperature is stated in degrees centigrade, it is an interval concept because the zero point is arbitrary. Temperature stated in degrees Kelvin is a ratio concept because the zero point is nonarbitrary. As will be clear from later discussion, the cephalic index is a comparative concept. If human skulls are classified as having a low, average, or high cephalic index, then the cephalic index is being used as a classificatory concept. There are concepts that can be conceptualized only at the level of classificatory concepts. Examples are metals and nonmetals; when animals and plants are divided into classes and further divided into orders, families, genera, and, finally species; when phenomena above the surface of the earth are divided into meteorological phenomena and astronomical phenomena.

Classes may exhibit order according to some underlying property. When objects are classified according to whether they are hot, warm, or cold, these are classified according to the five classes given earlier, there is an order to the five classes according to the size of the

cephalic index. Rating scales are very commonly used in psychology and education. The rater is required to classify behavior into two or more ordered classes of behavior. A very common form of rating scale requires the rater to indicate which of five classes his agreement-disagreement falls into, as the following example illustrates:

The instructor was very helpful to me.

A	B	C	D	E
Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree

Usually for the purposes of mathematical manipulation, the numbers five to one are attached to the categories A to E, respectively. Another example is a rating scale for classifying a student's behavior according to the extent that the student follows instructions:

A	B	C	D	E
Attends to written and oral instructions; follows directions accurately		Sometimes lets his attention wander; usually follows directions		Pays little attention to instructions; follows directions reluctantly or not at all

Here the categories B and D are behaviorally undefined. However, it is to be assumed that B contains behavior that falls in between A and C, and D contains behavior that falls in between C and E. It should be clear from these examples that a rating scale is a form of classifying behavior into an ordered series of classes.

The literature on classification does not always make is clear which is the classificatory concept--the concept referred to by the class name, or the criterion or criteria used for forming the classes. While

the class concept and class criteria are both properites, it is the class concept that is the classificatory concept. Thus in the example above relating to the classification of human skulls, each of the five classes represented the classificatory concepts and not the cephalic index which represented the criterion by which the classification was done. However, the meaning of each of the five classes cannot be understood without reference to the meaning of the cephalic index. Thus, while it is the class concept and not the criterion which is the classificatory concept, the class concept cannot be understood without reference to the criterion or criteria involved.

Changing a Classificatory Concept into a Comparative Concept

The reason for attempting to change a classificatory concept into a comparative concept is that measurement or some other way of assigning numbers may be used. That is, a person wants to change from the qualitative to the quantitative. Many tests are a way of changing a classificatory concept into a comparative one.

There are two ways to change a classificatory concept into a comparative concept, but there is no guarantee of the feasibility of either way. The first way is to conceptualize the classificatory concept in comparative terms, if possible. Carnap (1962, p. 12) describes the process as follows:

> The state of bodies with respect to heat can be described in the simplest and crudest way with the help of classificatory concepts like Hot, Warm, and Cold (and perhaps a few more). We may imagine an early, not recorded stage of the development of our language where only these classificatory terms were available. Later, an essential refinement of language took place by the introduction of a comparative term like 'warmer.' In the case of this example, as in many others, this second step was already made in the prescientific language. Finally,

> the corresponding quantitative* concept, that of temperature, was introduced in the construction of the scientific language.

Jones (1971, p. 340) makes a similar point as the following quote illustrates:

> Classification is, however, an essential prelude to measurement. The first steps towards progress in any science thus involve qualitative rather than quantitative distinctions. Properties must be recognized and classificatory procedures must be developed prior to the establishment of techniques for measurement. After a property has been successfully abstracted and similarity classes for that property have been defined, it may be discovered that successive classes lend themselves to quantitative interpretation; that is, the attribute-values of objects assigned to different classes may exhibit systematic differences in magnitude. If so, the quantitative definition of the attribute becomes possible, and measurement procedures may be devised. Measurement of an attribute, then, may evolve following successful efforts to generate a classification system.

The second way to change a classificatory concept into a comparative concept is somewhat artificial. It works with multidimensional classificatory concepts--that is, it works with those classifications where the inclusion of an object or phenomena is on the basis of multiple criteria. It also works with natural rather than artificial classifications. Hempel (1952, p. 53) explains the distinction:

> The rational core of the distinction between natural and artificial classifications is suggested by the consideration that in so-called natural classifications the determining characteristics are associated, universally or in a high percentage of all cases, with other characteristics, of which they are logically independent. Thus, the two groups of primary sex characteristics determining the division of humans into male and female are each associated--by general law or by statistical correlation--with a host of concomitant characteristics; this makes it psychologically quite understandable that the classification should have been viewed as one "really existing in nature"--as contrasted

*Carnap does not regard comparative concepts as quantitative.

> with an "artificial" division of humans according to the first letter in their given names, or even according to whether their weight does or does not exceed fifty pounds.

The classification of crystals is an example of a natural classification. The determining characteristics in the classification of crystals according to the number, relative length, and angles of inclination of their axes are expirically associated with a variety of other physical and chemical characteristics. Mendeleev's classification of elements in the periodic table is a natural classification which enabled him to predict the existence of several elements then missing in the table and to anticipate with great accuracy a number of their physical and chemical properties.

The taxonomic categories of genus, species, etc., as used in biology, determine classes whose elements share various biological characteristics other than those defining the classes in question. The classes may also reflect relations of phylogenetic descent. Thus the classes are chosen with a view to attaining systematic, and not merely descriptive, import. Mayr (1942, p. 10) states: "The devising of a classification is, to some extent, as practical a task as the identification of specimens, but at the same time it involves more speculation and theorizing." Gilmore (1940, p. 468) states:

> To sum up, . . . we are led to the view that a natural classification of living things is one which groups together individuals having a large number of attributes in common, whereas an artificial classification is composed of groups having only a small number of common attributes; further, that a natural classification can be used for a wide range of purposes, whereas an artificial classification is useful only for the limited purpose for which it was constructed; and lastly that both types are created by the classifier for the purpose of making inductive generalizations regarding living things.

As an example of how to change a classificatory concept into a comparative concept, the class of animals known as mammals, which is a

natural class, has been chosen. The members of the class have a large number of characteristics in common. Mammalinity, a comparative concept, has been created. Mammalinity is the degree to which the animal possesses characteristics which are typical of mammals. The following mammalinity inventory (in the form of a test) is a procedure for attaching numbers to indicate the degree of mammalinity. The items of the inventory were derived primarily from the work of Cochrum (1962, pp. 3, 4), and secondarily from Carrington (1963), Boorer (1971), DeBlase and Martin (1974), Morris (1965), and Hoffmeister (1963).

Mammalinity Inventory

Instructions: Each of the following items describes a characteristic which is typically possessed by mammals. Some of these characteristics are possessed by non-mammals which also belong to the phylum Chordata. All mammals will not possess each of the characteristics. The scale is suitable for use with any animal belonging to the phylum Chordata. To obtain a score for an animal, place a checkmark beside any characteristic it possesses; then count the number of checkmarks.

External Characteristics

1. Hair is present at some stage of the life cycle ___
2. An external ear opening, surrounded by a pinna (ear flap) ___
3. Sweat glands ___
4. Oil (sebaceous) glands ___

Internal Characteristics

I. Features of the soft anatomy

5. Mammary glands (milk glands) are present in females ___
6. The young are born alive from inside the animal ___
7. The brain has large cerebral hemispheres ___
8. A muscular diaphragm separates the lungs from the posterior body cavity ___

9. Well-developed facial muscles ___

10. The red blood cells are non-nucleated when fully developed ___

11. Warm-blooded ___

12. Muscles form a significant part of the body weight ___

13. A four-chambered heart ___

14. Large efficient (detection of minute concentations of chemicals in the air) noses. Sense of smell is most important sense of animal ___

15. Poor eyesight compared to birds ___

16. A valve, the epiglottis, at the opening to the windpipe ___

II. Osteological features

a. Skull

17. A double occipital condyle (formed by the exoccipal bones) ___

18. The zygomatic arch is an appendage of the skull instead of part of the skull ___

19. Each ramus of the mandible is composed of a single bone, the dentary ___

20. The jaw articulates directly with the squamosal ___

21. Three ear ossicles are present ___

22. The typanic bone surrounds and protects the inner ear ___

23. A secondary hard palate is formed from the premaxilla and maxilla ___

24. Teeth are present, different teeth having different specialized functions ___

b. Postcranial elements

25. Possess a backbone ___

26. Seven cervical or neck vertebrae ___

27. The limbs are rotated forward with marked angulation ___

28. The ankle joint is between the tibia and the tarsus ___

29. There are five metacarpal bones ___

Behavior

31. Female mammals care for their young ___

32. While growing into an adult, the animal indulges in play ___

The basic principle behind the Mammalinity Inventory is that the greater the number of mammalian characteristics an animal possesses, the greater is its mammalinity. As far as is known, the concept of mammalinity has no practical value and it has no theoretical import. The concept of mammalinity has little, if any, meaning. That is, it makes no sense to say that being a mammal exists in degrees. A sufficient condition for saying that an animal is a mammal is that it possesses milk-producing pores or glands. Once this characteristic has been used to identify an animal as a mammal, all other charactertistics are unnecessary for identification.

This second way of transforming a classificatory concept into a comparative concept does not require the concept to be conceptualized in comparative terms first. As illustrated by the Mammalinity Inventory, this procedure may result in a concept that has no real meaning in comparative terms. The procedure used for the Mammalinity Inventory is the very procedure that is used to turn educational and psychological classificatory concepts into comparative concepts in the form of tests. The danger of the procedure is that it may result in comparative concepts that are meaningless. Correspondingly the numbers that are attached by means of tests may be meaningless.

Comparative Concepts

Comparative concepts can be classified according to whether they are unidimensional or multidimensional. Unidimensional comparative concepts are weakly comparative.

Unidimensional Comparative Concepts. To establish a strong comparative concept of a unidimensional property for a given class or domain of objects or phenomena, D, is to specify criteria which determine for any two objects or instances of the phenomena in D whether they have the same amount of the property, and, if not, which of them has the smaller amount. By means of these criteria, it must be possible to arrange the elements of the given domain in a serial kind of order, in which an object or instance of the phenomena precedes another if it has a smaller amount of the property than another. Objects or phenomena of equal amounts of the property coincide, i.e., share the same place.

An example of a strong comparative concept is the mass of objects. It is possible to determine the comparative amounts of the mass of any two objects, x and y, by placing the objects in opposite pans of a beam balance. If x sinks and y rises, then y precedes x in mass. If x balances y, then the mass of x is said to coincide with the mass of y. To generalize, a comparative concept within the domain of application D is introduced by specifying criteria of coincidence and precedence for the elements of D in regard to the characteristic to be represented by the concept. The criteria of conicidence and precedence must be so chosen as to arrange the elements of D in a quasi-serial order, i.e., in an order that is serial except that several elements may coincide in order.

Multidimensional Comparative Concepts. Multidimensional comparative concepts are part of our everyday speech. The following are some examples:

> These two students are about equally intelligent.
>
> I would rank the College of Education at this University as being the fifth best in the nation.
>
> Peter knows more science than Mark.
>
> I am happier today than I was yesterday.
>
> Their ability at diving is much the same.
>
> This is a more beautiful painting than that one.
>
> His depression has become worse.
>
> Out of 10, I graded Bill's essay as 9 and John's essay as 8.

It should be obvious that each of the comparative concepts mentioned in the above example are multidimensional--intelligence, quality of a college of education, knowledge of science, happiness, ability at diving, beauty, and quality of an essay. In other words, each concept has a number of aspects, parts, or dimensions. For example, it is common practice to divide intelligence into verbal intelligence and numerical intelligence. Alternately, one could say that there are a large number of behaviors that could be classified as intelligent. On the whole, multidimensional concepts are usually less explicit in their meaning than are unidimensional concepts. For example, happiness and beauty are not highly explicit concepts and will vary with each person's perception of them. These concepts are often multidimensional comparative concepts.

In the above examples, comparisons of coincidence and precedence were given. Associated with the lack of explicitness of multidimensional comparative concepts is a lack of explicitness in the criteria for

coincidence and precedence. One may have trouble expressing in words what these criteria are--a case of tacit knowledge (Polany, 1958). Take one of the examples given above: "This is a more beautiful painting than that one." One may be able to specify some of the criteria for the stated precedence, but overall one may justify one's preference by saying "I just <u>feel</u> that this is a more beautiful painting than that one."

With unidimensional comparative concepts, one could state criteria of coincidence and precedence <u>in advance</u> of giving a comparative judgment. This is not always so with multidimensional comparative concepts. Returning to the example of the paintings, one may be able to give in advance only the most general criteria for the type of paintings one likes. However, only when one sees two paintings side by side may one be able to give more specific criteria as to why he prefers one to the other.

Numbers may be attached to multidimensional comparative concepts to indicate coincidence and precedence. There are three somewhat different ways of attaching numbers to multidimensional comparative concepts--ranking, rating, and indices. These three ways will not be discussed in detail in the next chapter, but a brief mention will be made of them here. "I would rank the College of Education at this University as being the fifth best in the nation" is an example of ranking. "Out of 10 points, I graded Bill's essay as 9 and John's essay as 8" is an example of rating. For a Grade 8 course in science, it might be decided that certain behavioral attributes are required. For each attribute (or a sample of attributes), a multiple-choice item is constructed. A score on the test is obtained by counting the number of attributes (i.e., items) that a student has. An index is an algebraic composite of a number of parts. Thus counting (i.e., adding) the number of items

correctly answered in a test results in an index. But not all indices are of this mathematical form. For example, the cephalic index, discussed in relation to classificatory concepts, is a percentage. The concept of the cephalic index is itself a multidimensional comparative concept.

The numbers attached to comparative concepts indicate coincidence and precedence. But the criteria of coincidence and precedence are weaker for multidimensional concepts than in the case of unidimensional concepts. I shall take tests as examples. Two students may obtain exactly the same score on a test, but it would be very unlikely that they answered all the same items correctly. They will have answered different items correctly, even though they both receive the same total score. So numerically their performances coincide but in actuality they may be quite different. If two objects balance each other on a beam balance, there is no doubt (within limits of error) that they have the same mass. To take another example, Bill may have obtained a score of four out of ten on a test, while John obtained a score of five. However, Bill may have answered correctly the four most difficult items while John answered the five easiest items. Irrespective of the numbers, Bill's performance is superior, or at least equal to, the performance of John. This is further illustration that multidimensional comparative concepts are weakly comparative.

Interval Concepts

Interval concepts are unidimensional concepts. They are an extension of unidimensional comparative concepts. As for comparative concepts, interval concepts also require criteria of coincidence and precedence. However, interval concepts in addition require a unit. When numbers are

assigned to a property conceived in comparative terms, one cannot tell how near to each other are different amounts of the property. As an example, consider three persons, A, B, and C, having IQs of 80, 100, and 120, respectively. Intelligence is a comparative concept and is not measured in units. We cannot tell from these numbers whether B is closer to A in intelligence or closer to C. We certainly cannot infer that B is equidistant from A and C in intelligence. By contrast, temperature can be conceived as an interval concept as it is when measured in units of degrees centigrade or degrees fahrenheit. Suppose we have three objects, A, B, and C, with temperatures of 80, 100 and 120 degrees fahrenheit, respectively. Then B can be said to be equidistant in temperature between A and C. That is, the temperature interval between A and B equals the temperature interval between B and C.

With an interval concept of a property, the assignment of numbers to the property of any two objects fixes the numbers for all the other objects. The initial assignment corresponds to selecting both an origin and a unit of measurement. Thus the centigrade scale of temperature is an interval scale: 0 and 100 are arbitrarily assigned to the freezing point and the boiling point of water. The temperature range between these is divided into a hundred equal intervals known as "degrees centigrade." The fahrenheit scale is also an interval scale, though with different origin and unit. Thirty-two degrees and 212 degrees are taken to be the freezing and boiling points of water, and zero degrees fahrenheit is the temperature of an equal mixture by weight of salt and snow.

Ratio Concepts

Ratio concepts are unidimensional concepts. They are extension of interval concepts of property. As for interval concepts, equal intervals of a property can be ascertained. The difference is that a zero point is no longer arbitrary. Length, mass, time, and current electricity are examples of ratio concepts. When temperature is measured in degrees Kelvin, it is conceived as ratio concept, zero degrees Kelvin being the lowest possible temperature, which occurs at minus 273 degrees centigrade.

As for interval concepts, the unit of measurement for ratio concepts is still arbitrary. Units are laid down by international agreement. For example, in 1960, the Eleventh General Conference of Weights and Measures, with 38 countries represented, sanctioned an international meter at 1,650,763.73 vacuum wavelengths of monochromatic orange light emitted by a krypton atom of mass 86. The number attached to a ratio property is not unique since it depends on the unit used.

Absolute Concepts

An absolute concept is different from a ratio concept in that there is no arbitrary choice of unit. Numerosity, the number of entities or attributes in a group, is an example of an absolute concept. When a ratio concept is measured, the resulting number is dependent on the choice of unit. However, the number of entities in a group can be determined uniquely. Counting (or enumeration) is the name given to the process by which numerosity is determined. In tests, the number of attributes is counted, i.e., the number of correct behavioral responses to a group of items is counted.

In summary, the difference between the qualitative and quantitative rests on the conceptualization that is involved. The nature

of the concept determines whether it can be quantified. The nature of the concept also determines the nature of the quantification. The following types of concepts were discussed: classificatory, comparative, interval, ratio, and absolute. Comparative, interval, ratio, and absolute concepts are known collectively as quantitative concepts, whereas classificatory concepts are qualitative concepts. Whereas entities, phenomena, and properties can all be classified, only properties can be quantified. A property that can be quantified is one for which it is possible to say that entity or phenomena A has more of the property, less of the property, or an equal amount of the property, compared to entity or phenomena B. A quantitative concept refers to a property that can be quantified.

PART II: THE WAYS OF ASSIGNING NUMBERS

The previous part emphasized the centrality of conceptualization to the process of quantification. The nature of a concept determines whether or not it can be quantified. Comparative concepts can be quantified but classificatory concepts cannot. If a concept can be quantified, its nature determines the way it can be quantified. Comparative, interval, ratio, and absolute concepts are all quantified in different ways. This part examines in detail different ways of quantifying. However, this part is broader. It looks at the different ways of assigning numbers, not all of which are forms of quantification (e.g., naming by numbers is not a form of quantification).

The most common definition of measurement that is given in the literature is that measurement is the assignment of numbers to the properties of objects and events according to rules. It is the contention of this part that measurement is only one of at least six fairly distinct ways of assigning numbers.

1. Measurement
2. Counting
3. Naming by Numbers
4. Numerical Judgment
5. Indices
6. Ranking

Beginning with measurement, each of the six different ways will be discussed. By beginning with a description of measurement, it should be clear how the other five ways are quite different and should not be confused with measurement. Once the differences among these ways of assigning numbers are understood, it becomes clear that tests are not a form of measurement, as measurement is understood in the physical and biological sciences.

Measurement

The first four sections of this discussion of measurement detail the nature of measurement and lead to a definition of measurement. The first section gives an introduction to measurement using the notion of sets. The next section explains what fundamental measurement is and indicates that it is fundamental measurement that is being discussed in this part. The third section discusses the three basic properties for determining the empirical relations between quantities of a property. The fourth section deals with the representation problem and gives a formal definition of measurement. The representation problem is concerned with the justification of the assignment of numbers to objects or phenomena.

The fifth section gives the many differing definitions of measurement found in the literature. These definitions are compared to the one developed in this part. The popular definition given by Stevens (1951:22) is analyzed for its deficiencies.

The sixth section introduces the steps that lead to the measurement of a property. This is done by looking at an historical example in the physical sciences.

The final section answers the question of why we measure by looking at how measurements are used. Constant reference is made to tests and test theory since tests are commonly used in evaluation. This section shows that test theory has not explicitly examined the evolutionary steps that lead to the measurement of a property and instead has concentrated on the various aspects of using measurements.

Introduction to Measurement Via Sets

One way of understanding measurement is through the notion of a set. A set is a collection of objects or elements. Measurement can be thought

of as involving two sets--a set of objects or events and a set of numbers. Figure 1 is a diagrammatic representation of the measurement process. Set A represents a set of objects x, y, and z, whose properties are to be measured. The objects x, y, and z have been assigned the numbers 10, 4, and 6 respectively, which are denoted symbolically as f(x), f(y), and f(z). This process of assigning numbers is called mapping. It is said that the members of one set are mapped onto the members of another set by means of a rule of correspondence. Another name for a rule of correspondence is a function, symbolized as f.

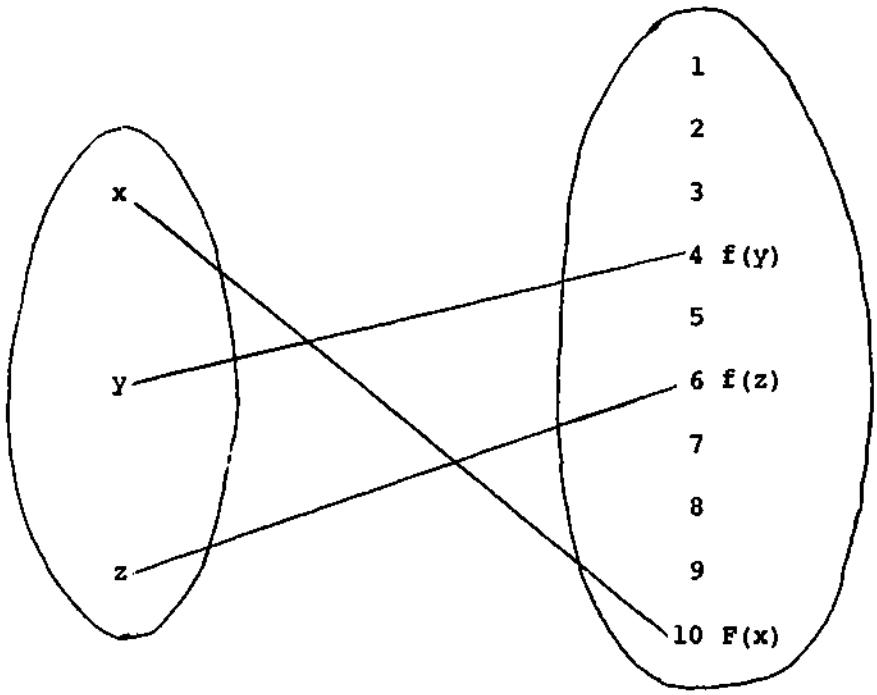

Figure 1. A diagrammatic representation of the measurement process.

Set Re is said to be a homomorphic image of set A if the relations that exist between x, y, and z also exist between f(x), f(y), and f(z). For example, suppose it can be determined empirically, independently of

the measurement process, that x is greater than z and z is greather than y, i.e., $x > z$ and $z > y$. The numbers that have been assigned reflect these relations, since $10 > 6$ and $6 > 4$, i.e., $f(x) > f(z)$ and $f(z) > f(y)$. Furthermore, suppose it can be determined independently of the measuring procedure that $x = y + z$. This relation also holds between the numbers, since $10 = 4 + 6$, i.e., $f(x) = f(y) + f(z)$. It should be clear that the rules for assignment of numbers must be chosen with care. The rules must be such that the relations that hold between the magnitudes of the properties of the objects must also hold between the numbers that have been assigned.

Fundamental Measurement

The measurement of the density of a cube depends on the measurement of mass and volume. The mass can be measured directly by means of a beam balance. The volume of the cube can be obtained as a result of measuring the length of the side of the cube with a scale. In this example, density and volume were measured indirectly by measuring other quantities. This was not true for mass and length which were measured directly. Measurement which is direct is known as fundamental measurement*--that is, a property is said to have a fundamental measure when no prior measurement of other quantities is required (Krantz, Luce, Suppes, and Tversky, 1971:1).

Fundamental measurement does not apply only to the fundamental quantities of physics, which include length, mass, time, and the flow of

*Different authors use different terminology. I shall use that of Krantz, Luce, Suppes, and Tversky (1971). Ellis (1966:55-56) used the term "direct" for fundamental measurement and "fundamental" to refer to extensive measurement. Extensive measurement involves those attributes which can be added. Campbell (1920:267-294) uses the term "fundamental" to refer to both fundamental and extensive measurement.

electricity. For example, the volume of liquids can be measured fundamentally by means of a measuring cylinder. Electrical resistance can be measured fundamentally using a Wheatstone's Bridge apparatus.

All of Part II is concerned with fundamental measurement. Fundamental measurement is the most basic kind of measurement. Indirect measurement inevitably depends on the fundamental measurement of other quantities. In the density example above, the measurement of density depended on the fundamental measurement of mass and length. In the social sciences, indirect measurement often involves measuring a fundamental quantity from physics. For example, rate of learning may be measured by the length of time required to learn a list of nonsense syllables. On the whole, the social sciences have not developed the fundamental measurement of social science concepts. Thus the examples of measurement given in the discussion of measurement are largely drawn from the physical sciences.

Basic Procedures for Determining the Empirical Relationships Between Quantities of a Property

In the section introducing measurement via sets, it was supposed that empirical relationships could be determined independently of the measuring process. This section examines the three basic procedures for determining empirical relationships (Krantz, Luce, Suppes, and Tversky, 1971). Emphasis will be placed on the first two, since the occurrence of the third is not so common.

The first procedure is the ordinal procedure. It is concerned with the arrangement of a property p in order according to size. The relationships "greater in p than," "equal in p to," and "less in p than" are determinable independently of any measuring procedure (Ellis, 1966:74). In the case of length, the procedure is to place pairs of

objects side by side lengthwise. Thus if we place two rigid straight rods x and y side by side and adjust them so that one is entirely beside the other and they coincide at one end, then either x extends beyond y at the other end, or y extends beyond x, or they appear to coincide at that end. We say, respectively, that x is longer than y, y is longer than x, or that x and y are equivalent in length. Symbolically, we write, respectively, $x \succ y$, $y \succ x$, or $x \sim y$. In the case of mass, pairs of objects can be compared by using a beam balance. An object is placed in each of the two pans. We observe which pan descends, and this will indicate which object is heavier. If there is balance, then the two objects are equivalent in mass. A numerical assignment can be checked by means of this ordinal procedure. If $f(x) > f(y)$ or $f(y) > f(x)$ or $f(x) = f(y)$, then $x \succ y$, $y \succ x$, and $x \sim y$ respectively. Properties for which this ordinal procedure cannot be carried out cannot be measured. Examples of such properties are odor, gaseous, quadruped, reversibility, and diatomic.

The second procedure is concerned with the counting of units using standard sequences and involves the concatenation of objects so that the property of interest is combined. For example, two or more straight rigid rods can be concatenated by laying them end to end in a straight line, and so we can qualitatively compare the length of one set of concatenated rods with that of another by placing them side by side, just as with single rods. The concatenation of x and y is denoted by $x \circ y$, and the observation that z is equal to $x \circ y$ is denoted $z \sim x \circ y$. We can concatenate the mass of two objects by placing them together in a scale pan so that their combined mass may be compared with a third object.

There are many properties which cannot be combined. Examples are temperature, density, hardness, and most social science concepts. Two volumes of liquid, both at the same temperature, when combined into one

body of liquid, will not have a temperature greater than that of the original two. In fact, there will be no change in temperature. In the same way two volumes of liquid, all of the same density, when combined, will not have a density any different from the original two volumes. Combining two people of equal intelligence to form a group, will not make the group twice as intelligent as any one of its members. Properties like temperature, density, hardness, and intelligence are said to be "intensive" whereas properties that can be combined are said to be "extensive." It is possible that extensive properties may be discovered in the social sciences.

On the basis of ordering and concatenation it is possible to set up a scale for the measurement of a property. Suppose that $\underline{x}'$, $\underline{x}''$, $\underline{x}'''$, ... are perfect copies of $\underline{x}$ with regard to the property under consideration. By the procedure of concatenation it is possible to construct what is called a standard sequence, i.e.,

$$x,$$

$$2x \backsim x \circ x',$$

$$3x \backsim 2x \circ x'',$$

$$4x \backsim 3x \circ x''',$$

and so on.

In the case of length, a meter stick graded in millimeters provides, in concenient form, the first 1,000 members of a standard sequence constructed from a one-millimeter rod. With the first procedure of ordering and with the second procedure of counting using the standard sequence, it is possible to assign a number to the length of a rod $\underline{y}$. Rod $\underline{y}$ is placed beside the meter stick so that they coincide at one end. If we observe that rod $\underline{y}$ falls between $\underline{nx}$ and $(n + 1)x$, say, between 325 and 326 mm, then we assign it a length between $nf(x)$ and $(n + 1)f(x)$ (in

the present example, between 325 f(x) and 326 f(x) where f(x) is the numnber assigned to a one-millimeter rod and its copies). The value of f(x) depends on the selection of a particular rod, u, to have unit length. If $u \sim mx$, then f(x) = 1/m. Thus if the unit length, u, is the meter, then m = 1,000 and the length assigned to y must be between 0.325 and 0.326 meters. If the unit of length is the centimeter, then m = 10 and f(y) must be between 32.5 and 32.6 centimeters.

Application of ordinary arithmetic to the numbers assigned is not meaningfully possible when only the ordinal procedure (i.e., the first procedure) can be carried out with a property. For example, if a child is ranked second in class and another child is ranked seventh, it is not meaningful to add two and seven. While addition is not meaningful, neither is subtraction, multiplication, or division. However, if a property can be concatenated and numbers are assigned using a standard sequence, then arithmetical operations can be carried out. As Hanson (1969:50) emphasizes, what arithmetical relations hold between the numbers must be checked out by experimentally determining what relations hold between the magnitudes of the properties.

> Thus, suppose that object A is regarded as having the unit weight. We can assign weights to other objects by the process described, such that A_2 will have weight 2, A_4 will have weight 4 and A_6 weight 6. Now, can we be certain, in advance of experiment that A_2 and A_4 will, if placed together in one pair of the beam balance, just balance A_6 placed in the other? No. It is very important to note that we cannot be certain of this until we actually perform the experiment. The emphasis of some scientists, and the usual exaggeration of that emphasis by philosophers, has sometimes been such as to minimize this. That 2 + 4 = 6 can be demonstrated in pure arithmetic without experiment--that's a matter of formal prescription. But until we performed the proper experiments we could not be sure that the physical operation of addition of weights would exhibit the familiar properties of purely arithmetical addition--for physical properties are a matter of how nature is put together.

Displacement, velocity, acceleration, and force are examples of vector quantities to which ordinary (scalar) arithmetic does not apply. As an example, suppose a force of one newton and a force of two newtons act on a body, the angle between the two forces being 120 degrees. The combined or resultant force is not three newtons, but $\sqrt{3}$ newtons in a direction at right angles to the force of one newton. This illustrates that what arithmetic can be applied to the assigned numbers must be checked empirically.

For the sake of completeness, the third basic procedure for determining the empirical relationships between quantities of a property will be mentioned. It involves solving inequalities. krantz, Luce, Suppes, and Tversky (1971:5) describe the procedure as follows:

> Suppose that five rods denoted $a_1, a_2, \ldots a_5$, are found to satisfy
>
> $a_1 \circ a_5 \quad a_3 \circ a_4 \succ a_1 \circ a_2 \succ a_5 \succ a_4 \succ a_3 \succ a_2 \succ a_1$
>
> Data such as these can arise whenever a limited set of preselected objects and concatenations are compared and where it is impractical to go through the elaborate process of constructing standard sequences. Denote by x_i the unknown value of the length of a_i . . . . From the above observations, the unknown lengths x_i must satisfy the following system of simultaneous linear inequalities:
>
> $$\begin{aligned} x_1 + x_5 - x_3 - x_4 &> 0, \\ x_3 + x_4 - x_1 - x_2 &> 0, \\ x_1 + x_2 - x_5 &> 0, \\ x_5 - x_4 &> 0, \\ x_4 - x_3 &> 0, \\ x_3 - x_2 &> 0, \\ x_2 - x_1 &> 0. \end{aligned}$$
>
> Any solution to this set of seven inequalities in five unknowns gives a possible set of values for the lengths of $a_1, \ldots, a_5$. One can thus measure the five rods by finding a solution, if one exists.

The Representation Problem

This section begins a more formal discussion of measurement than that given in the previous sections. The representation problem is concerned with the justification of the assignment of numbers to objects or phenomena (Suppes and Zinnes, 1963:4).

Whenever we measure a property, a numerical model of the world is constructed. A numerical model is regarded as a model of the world if it reflects the structure of the world or presents its essential features. To make the representation problem more precise, we introduce the notion of a relational structure which leads to a way of characterizing the nature of the correspondence between the world and its numerical model.

A relational structure is a collection of objects, phenomena, or numbers along with one or more relations defined among them. Formally, a relational structure is a sequence $\langle A, R_1, \ldots, R_n \rangle$ where A is a non-empty set and $R_1, \ldots, R_n$ are relations defined on the elements of A. Angle brackets, $\langle\ \rangle$, rather than parentheses are used in giving an explicit listing of a relational structure. A relational structure is said to be empirical if A contains objects or phenomena and it is said to be numerical if A contains numbers.

Consider a simple relational structure of the form $\langle A, R \rangle$ where R is a binary relation, that is, a relation between pairs of entities in A. For example, A could be a set of objects and R could be the relation "longer than." For any pair of objects $\underline{x}$, $\underline{y}$ in A, we define:

x R y if and only if $\underline{x}$ is longer than $\underline{y}$

An example of a numerical relational structure would be the case where A is the set of all real numbers and R is the relation "greater than." In this case, if $\underline{x}$ and $\underline{y}$ represent real numbers, then:

x R y if and only if $x > y$.

Measurement can be described as the representation of an empirical (relational) structure by a numerical (relational) one (Coombs, Dawes, and Tversky, 1970:10). In formal terms, an empirical structure $\mathcal{A} = \langle A, R \rangle$ is said to be represented by a numerical system $\mathcal{B} = \langle Re, S \rangle$ * if there exists a function f from A into Re (which assigns to each x in A an f(x) in Re) such that for all x, y in A,

x R y implies f(x)Sf(y).

Thus $\mathcal{A}$ is represented by $\mathcal{B}$ if there exists a correspondence f that maps A into Re in such a way that if the relation R holds between some x and y in A, then the relation S holds between f(x) and f(y) in Re, where f(x) and f(y) are the images of x and y, respectively. If the function f is a one-to-one correspondence (i.e., f assigns to each x in A a unique f(x) in Re), then $\mathcal{A}$ and $\mathcal{B}$ are said to be isomorphic. However, in practice it is too strict to require that the function f be one-to-one, for it may be necessary to assign the same number to two distinct objects, as when two objects have the same length or weight, for example. We then say that $\mathcal{B}$ is a homomorphic image of $\mathcal{A}$. Measurement can be defined as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful (Krantz, Luce, Suppes, and Tversky, 1971:9).

The measurement of length discussed in the previous section is a good example of the definition. In terms of the representation problem, let $\mathcal{A} = \langle A, \succ, \circ \rangle$ where A denotes a set of empirical objects, $\succ$ denotes the relation "longer than," and $\circ$ denotes a physical concatenation. The concatenation operation represents a ternary relation on A, holding among

---

*Re refers to the set of real numbers; S is a binary relation between pairs of numbers.

x, y, and $z \sim x \circ y$, whereas $\succ$ is a binary relation on A. Let $\mathcal{B} = \langle$Re, $>$, + $\rangle$ where Re denotes the positive real numbers, $>$ denotes the usual inequality between real numbers, and + denotes scalar addition between real numbers. The measurement of length is essentially a representation of $\mathcal{A}$ by $\mathcal{B}$. It consists of mapping A into Re in such a way that the number assigned to one object is greater than that assigned to a second object whenever the first object is longer than the second, and such that the number assigned to the concatenation of two objects equals the sum of the numbers assigned to the two separate objects. The numerical assignment f is a homomorphism in the sense that it sends A into Re, $\succ$ into $>$, and $\circ$ into + in such a way that $>$ preserves the properties of $\succ$ and + the properties of $\circ$.

After the representation problem has been solved and the scale is constructed, the next problem to be solved is the uniqueness problem. The uniqueness problem poses the question: given a particular measurement procedure, how much freedom is there in assigning numbers to objects or events? Answering this question leads to a discussion of scale types (i.e., ordinal, interval, ration, etc.). Since such discussions are commonly found in the literature, they will not be dealt with here.

To summarize Part II thus far, the discussion has led to an explanation of measurement in terms of relational structures. Measurement can be defined as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful. That is, we will know that we have a scale of measurements when the numerical structure or model represents the empirical structure or model. To check on this representation, the relations that hold for the empirical structure must

be determinable independently of the measuring procedure. For example, if $f(x) \succ f(y)$, $f(y) \succ f(z)$, and $f(x) \succ f(z)$, then it must be shown independently of the measuring procedure, that $x \succ y$, $y \succ z$, and $x \succ z$.

The Various Definitions of Measurement

This section presents the different definitions of measurement given by a number of authors. These definitions are compared with the definition of measurement developed in the previous section. The popular definition of measurement given by Stevens (1951:22) is analyzed and criticized.

In the preceding sections of this chapter, the view of measurement subscribed to is that given by Krantz, Luce, Suppes, and Tversky (1971:9). To them, ". . . measurement may be regarded as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful." What makes this definition more restrictive than others is that it requires the existence of an empirical relational structure independent of the numerical relational structure.

Torgerson (1958:14) presents a similar view in less precise but simpler language:

> Measurement of a property then involves the assignment of numbers to systems to represent that property. In order to represent the property, an isomorphism, i.e., a one-to-one relationship, must obtain between certain characteristics of the number system involved and the relations between various quantities (instances) of the property to be measured.
>
> The essence of the procedure is the assignment of numbers in such a way as to reflect this one-to-one correspondence between these characteristics of the numbers and the corresponding relations between the quantities.

Lord and Novick (1968:17) and Stevens (1959:20) give definitions similar to the previous two. Lord and Novick state:

> We shall define measurement to mean a procedure for the assignment of numbers (scores, measurements) to specified properties of experimental units in such a way as to characterize and preserve specified relationships in the behavioral domain.

They refer to the behavioral domain since this definition is taken from a book on the statistical theories of mental test scores. Stevens states his view of measurement as follows:

> Under the modern view, the process of measurement is the process of mapping empirical properties or relations into a formal model. Measurement is possible only because there is a kind of isomorphism between (1) the empirical relations among properties of objects and events and (2) the properties of the formal game in which numerals are the pawns and operators the moves.

Elsewhere in the same paper, Stevens (1959:18) takes a much less restrictive view of measurement:

> It is no new thing, of course, to find practice outrunning legislation, for that is the nub of the story of mathematics. The irrationals, the surds, the imaginaries, and the negatives are numbers that still bear names reminiscent of protest--protest against outlandish practice and against the writing of unauthorized absurdities. But orthodoxy bent to accommodate practice. Mathematicians staved off chaos by rationalizing the use of irrationals, and by imagining a broader domain in which imaginaries and negatives could serve as proper elements. An analogous story can be told of measurement. The reach of this concept is becoming enlarged to include as measurement the assignment of numerals to objects or events according to rule--any rule.

As shall shortly be discussed, this view is too unrestrictive, because not just any rule will do. Other unrestrictive definitions of measurement are as follows. Nagel (1960:121) states:

> Measurement has been defined as the correlation with numbers of entities which are not numbers

Siegel (1956:29) states:

> Measurement is the process of mapping or assigning numbers to objects or observations.

The most commonly reported definition is that of Stevens (1951:22):

> Measurement is the assignment of numerals to objects or events according to rules.

Ellis (1966:39) objects strongly to the notion that any rule will do. He illustrates his point with an example. Suppose that on a table there are a child's tractor, an empty coffee cup, an ink bottle, and an empty packet of cigarettes. Now suppose that a person is instructed to take these various objects in turn and assign to each a number--the first that comes into the person's head. Suppose that the numbers actually assigned are 2, 2, 2, 3. There is no doubt that the person has followed a rule, which was to assign the first numbers that came into his head. However, it is doubtful whether anyone would say that measurement had been carried out. One could easily think of other rules that would not constitute measurement, such as assign numerals according to the throws of a die, or assign telephone numbers according to the order in which they are found in the telephone book. Ellis (1966:40) gives another example of what he considers as the inadequacy of just any rules:

> Again, suppose that I am instructed to take any monotonic increasing sequence of rational numbers, and assign the first number to the first book on the shelf in front of me, the second number to the second book, and so on. Can I then be said to have measured the books in any way? Clearly, the numerical assignments have been made according to a rule. But how should I express the results of my "measurements"? Suppose that the second book is A Textbook of General Botany and that the second number in my sequence is 37.53. Should I now say that A Textbook of General Botany is 37.53? If so, what information does this statement carry--even to someone who knows the rule which led to this numerical assignment? It does not tell him that it is the second book on the shelf in front of me. For he has no way of knowing what sequence of rational numbers I chose. Moreover, there is no question

> of his being able to check the "measurement." If he too follows the rule, and makes a numerical assignment to A Textbook of General Botany, it is extremely unlikely that he will make the same numerical assignment. But clearly his "measurement" would not conflict with mine.

Ellis believes that a rule must lead to a scale of measurement satisfying two conditions. First, different measurements made on the same scale on the same particular under the same conditions should not conflict with one another. Second, when measurements are made on a particular scale, the statements of the results of these measurements must be informative. Thus Ellis (1966:41) modifies Stevens' definition of measurement to read:

> (a) Measurement is the assignment of numerals to things according to any determinative, non-degenerate rule.
> (b) We hve a scale of measurement if and only if we have such a rule.

By determinative, Ellis means that the same numerals would always be assigned to the same things under the same conditions. A non-degenerate rule allows for the possibility of assigning different numerals to different things, or to the same thing under different conditions. A degenerate rule would be: "Assign the number 3 to everything." According to Ellis, the numerical assignments made according to a determinative, non-degenerate rule will be informative.

Ellis' criticism of Stevens' definition that not just any rule will do seems justified. However, it is possible to specify more clearly than Ellis what is necessary for a rule to be informative. The purpose of measurement is to inform one about the empirical relations that exist between the various magnitudes of the quantitative property being measured. In other words, the numerical relational structure that results from the application of the rule must be such that it is homomorphic with the empirical relational structure of interest.

Stevens' definition describes measurement as the assignment of numerals to objects or events. This is too unrestrictive. Measurement involves the assignment of numberals to the properties of objects or events. To Russell (1938:176), "Measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real as the case may be." To Campbell (1938:126), measurement is "the assignment of numerals to represent properties of material systems other than number, in virtue of the laws governing these properties." For Russell, numbers correspond to "magnitudes" and for Campbell, they represent "properties of material systems." However for Stevens, numbers are assigned "to objects or events." By "magnitude" Russell means an amount of a property and thus Russell is in agreement with Campbell on this point. Stevens' definition does not mention property. For him, if numerals are assigned to objects according to rules, we have measurement. Apparently it is the object that is measured, and not (at least not necessarily) a property of the object. Torgerson (1958:14) states:

> [Stevens] does not object to the use of the term measurement to denote, say, the sorting of sticks into piles according to whether they grow on oak, elm, or pine trees--as long as numerals are used for naming the piles rather than words. According to this view, we have measured or scaled a stick, though only at a primitive, nominal level, when we determine that that particular stick is a "two." Thus, for this approach, classification, or even naming of individual instances, becomes a kind of measurement.
>
> We shall not use the term measurement in this way. We shall rather retain the more traditional view, that measurement pertains to properties of objects, and not to the objects themselves. Thus, a stick is not measurable in our use of the term, although its length, weight, diameter, and hardness might well be.

Cohen and Nagel (1934:294) and Campbell (1938:122) put forward views similar to those of Torgerson. Campbell argues, "A street is not measured when numerals are assigned to the houses in it; a dyer does not measure his colours when he assigns numbers to them in his catalogue."

Siegel (1956:22) goes so far as to suggest that the attachment of names to classifications is a form of measurement:

> Measurement at its weakest level exists when numbers or other symbols are used simply to classify an object, person, or characteristic. When numbers or other symbols are used to identify the groups to which various objects belong, these numbers or symbols constitute a nominal or classificatory scale. . . . The psychiatric system of diagnostic groups constitutes a nominal scale. When a diagnostician identifies a person as "schizophrenic," "paranoid," "manic-depressive," or "psychoneurotic," he is using a symbol to represent the class of persons to which this person belongs, and thus he is using nominal scaling.

So, under a nominal scale of measurement, Siegel would include the case where names are attached to classificatory categories. Both Siegel and Stevens appear to view classification as a form of measurement. In Part I classification was viewed as a form of conceptualization. The attachment of numbers or names to categories is a labelling process quite distinct from the way in which I have described the process of measurement.

In summary, some authors present a definition of measurement which is very similar to one developed in the previous section. Stevens gives a very unrestricted definition of measurement. Ellis criticizes Stevens since Ellis believes that the assignment of numerals cannot be according to just <u>any</u> rule. Stevens does not specify that the assignment of numerals is to the <u>property</u> of objects or events. Since he does not, this allows the labelling of classes with numbers to be considered a form of measurement. However, naming by numbers is quite distinct from measurement.

The Steps that Lead to the Measurement of a Property-- An Historical Example

It is the view of this section that there are four steps that lead to the measurement of a property.

1. Qualitative observation is required in order to understand the property.

2. The property must be conceptualized in quantitative terms.

3. There must be the development of a procedure for determining the empirical relations between amounts of the property.

4. A way of assigning numbers must be devised so that the resulting numerical relational structure is homomorphic with the empirical relational structure.

These four steps will be illustrated by discussing an historical example. A brief history of the development of some aspects of kinematics based on the writings of Toulmin and Goodfield (1961) will be given. Kinematics deals with the movements of bodies in terms of distance, time, velocity (speed), and acceleration. It is not concerned with the forces and causes responsible, which form the subject-matter of dynamics. However, in order to understand the relationship between force and motion, it was necessary first to develop the concept of acceleration which leads especially to Newton's second law of motion, that force is equal to mass times acceleration. Velocity is the rate of change of position over time. We commonly talk about velocity as the miles per hour shown by the speedometer of a car. Velocity can be calculated by dividing the change in position by the time taken. Acceleration is the rate of change of velocity over time. For example, in free fall in a vacuum at the surface of the earth, a body moves with an acceleration of approximately 32 feet per second per second. That is, in every second of its motion, the velocity increases by 32 feet per second. As conceived from the seventeenth century, velocity and acceleration are quantitative

concepts, and there are mathematical relations among velocity, acceleration, distance, and time.

The first major figure in the study of kinematics was Aristotle. He was a brilliant zoologist who was impressed with the complexity, variety, and vitality of Nature. Consequently, he was never convinced that one either could or should reduce the workings of Nature to abstract, mathematical terms. He was interested in movement as a qualitative pehnomenon to be explained in the same sort of terms as are changes in color, warmth, or health. In his discussion of movement, there is only a minimum of quantification, such as simple numerical ratios, for example, between one distance and another. A mathematical concept such as velocity which is not a simple number but rather distance divided by time, led to difficulty for him. How can you divide a length by a time and get a "pure" ratio, he argued.

Like other Greeks, Aristotle used the words "faster" and "slower," and always specified velocity in terms of actual distances travelled in given times. Objects do not have a velocity of x m.p.h. but they are shifted y miles in z hours. Even Archimedes expressed his kinematic theories in the same terms: "If some point is moved with a uniform velocity along [the whole length of] a given line, and if we mark out upon this like two [shorter] lines, these will bear the same ratio to one another [in length] as do the periods of time taken by the point in traversing them" (Toulmin and Goodfield, 1961:213).

The point that I want to make is that the ancient Greeks seemed to be asking themselves basic questions: Is the property essentially a qualitative or a quantitative one? If the property is quantified in a certain way, is this an adequate way to quantify it? In relation to test concepts, these basic questions are rarely--if ever--asked. Just because

numbers have been assigned to a property does not mean that quantification has taken place. Numbers may be assigned to properties that are essentially qualitative ones.

It was the mediaeval mathematicians who recognized that a body's velocity could be treated as a quantitative variable in its own right--not just as a distance gone in a standard time. Their next problem was to describe the motion of an accelerating body in terms of this property, velocity, which might change continuously from one instant to the next. One important development was that of graphical techniques to show logically the relationship among uniformly changing velocity, distance, and time. (For these graphical techniques, see Toulmin and Goodfield, 1961:215.) These graphical procedures encouraged them to look for ways of replacing qualities, which Aristotle had regarded as fundamental, by numerical degrees of quantities. Mediaeval scholars such as Heytesbury and Oresme were able to prove that, if a body were ever to accelerate from rest uniformly, then its distance from the starting point must by definition increase in proportion to the square of the time. It was Galileo, with the use of measurements, who gave experimental demonstration in the real world, of these logical, abstract ideas. This demonstration is described in a famous passage in his Discourses on Two New Sciences:

> A piece of wooden moulding or scantling, about 12 cubits long, half a cubit wide, and three fingerbreadths thick, was taken; on its edge was cut out a channel a little more than one finger in breadth; having made this groove very straight, smooth, and polished, and having lined it with parchment, also as smooth and polished as possible, we rolled along it a hard, smooth, and very round bronze ball. Having placed this board in a sloping position, by lifting one end some one or two cubits above the other, we rolled the ball, as I was just saying, along the channel, noting, in a manner presently to be described, the time required to make the descent. We repeated this experiment more than once in order to measure the time with an

> accuracy such that the deviation between two observations never exceeded one-tenth of a pulse beat. Having performed this operation and having assured ourselves of its reliability, we now rolled the ball only one-quarter the length of the channel; and having measured the time of its descent, we found it precisely one-half of the former. Next we tried other distances, comparing the time for the whole length with that for the half, or with that for two-thirds, or three-fourths, or indeed for any fraction; in such experiments, repeated a full hundred times, we always found that the spaces traversed were to each other as the squares of the times, and this was true for all inclinations of the plane, i.e., of the channel, along which we rolled the ball. We also observed that the times of descent, for various inclinations of the plane, bore to one another precisely that ratio which, as we shall see later, the Author had predicted and demonstrated for them.
>
> For the measurement of time, we employed a large vessel of water placed in an elevated position; to the bottom of this vessel was soldered a pipe of small diameter giving a thin jet of water, which we collected in a small glass during the time of each descent, whether for the whole length of the channel or for a part of its length; the water thus collected was weighed, after each descent, on a very accurate balance; the differences and ratios of these weights gave us the differences and ratios of the times, and this was with such accuracy that although the operation was repeated many, many times, there was no appreciable discrepancy in the results. (Toulmin and Goodfield, 1961:219).

In the measurement of time, Galileo uses the fact that the time of descent ($\underline{t}$) is proportional to the weight of the water collected ($\underline{w}$). Thus if the experiment is carried out on two different occasions, one and two, for two different distances (d), then $\frac{t_1}{t_2} = \frac{w_1}{w_2}$ which follows mathematically from the fact that the time is proportional to the weight. What is of concern in the experiment is the relations between the amounts of time. The relations between the amounts of times are equal to the relations between the measured weights of water. Time is not measured in the sense of assigning numbers, but it is the empirical relational structure that is determined by measuring the weight of water collected. A procedure for determining the empirical relational

structure is essential, because without it there cannot be a numerical relational structure. What philosophical discussions of measurement ignore is the importance of the procedure or instrument for determining the empirical relational structure. It is the assignment of the numbers that is the focus of the attention. Of course most instruments of measurement serve both functions--the determination of the empirical relational structure and the assignment of numbers to represent the structure.

To return to Galileo's experimental demonstration, as for time, it is the determination of the relative distances, and not the actual distances, that is important. To demonstrate that distance travelled is proportional to the square of the time, Galileo showed that

$$\frac{d_1}{d_2} = \left(\frac{t_1}{t_2}\right)^2 = \left(\frac{w_1}{w_2}\right)^2 .$$

To quote Galileo, ". . . we now rolled the ball only one-quarter the length of the channel; and having measured the time of descent, we found it precisely one-half of the former." That is,

$$\frac{d_1}{d_2} = \frac{1}{4}, \quad \frac{t_1}{t_2} = \frac{1}{2}, \quad \text{and} \quad \frac{1}{4} = \left(\frac{1}{2}\right)^2.$$

To summarize, the steps in the process of measurement of a property are as follows. These steps are not necessarily distinct from one another but may merge.

1. Qualitative observation is required in order to understand the property.

2. If the property is to be measured, it must be conceived in quantitative terms. The property may be found to be a qualitative one, and quantification is not then possible.

3. An instrument or procedure has to be devised for determining the empirical relations between amounts of the property.

4. A way of assigning numbers must be devised so that the resulting numerical relational structure is homomorphic with the empirical relational structure.

Using Measurements

This section answers the question of why we measure by looking at how measurements are used. Six broad questions that may be asked in relation to the use of measurement are identified and discussed. This section makes constant reference to test theory. The conclusion is that test theory has ignored the four steps (discussed in the previous section) that lead to the measurement of a property and instead has concentrated on the various aspects of using measurements.

Within science and technology, measurement has high prestige. The social sciences have attempted to capture some of this prestige by assigning numbers wherever possible. Why should measuring have this preferential status? What is it that measuring accomplishes that nonmeasuring does not? One answer is that quantitative information can be more precise. There is no reason to be precise for precision's sake, of course. Precise information is information that enables one to distinguish objects, events, phenomena, and their properties to some arbitrarily assigned degree of refinement. However, I do not want to imply that qualitative information does not allow one to make distinctions. One can distinguish between a shark and a whale, by saying the former belongs to the fishes and the latter mammals, the essential distinction lying in their reproductive systems. However, there are certain distinctions that we want to make that require quantification. For example, for building human shelters, primitive man found human judgment of length sufficiently precise. However, today we could not build our shelters without using measurements to make precise

distinctions about the length of objects. We make quantitative distinctions for a purpose and a use. Churchman (1959:84) proposes that "the function of measurement is to develop a method for generating a class of information that will be useful in a wide variety of problems and situations." The following is a list of seven broad uses of measurement:

1. Measurement has an important function in technology, crafts of various kinds, and in practical affairs (e.g., carpentry, cooking, bookkeeping).

2. In the application of established theories and quantitative generalizations, measurements are made for insertion into these theories and generalizations.

3. Measurement is used in the process of refining established theories so that the agreement between measurements predicted by theories and actual measurements is greater.

4. Measurement is carried out in the process of elaborating an established theory in order to increase its areas of applicability.

   The last three functions are all related to what Kuhn (1970) has called normal science. The next three functions are related to what Kuhn (1970) calls revolutionary science:

5. Measurement can aid in the choice between competing theories.

6. Measurement can display serious anomaly between measurements predicted by a theory and actual measures.

7. Measurement can aid in the confirmation of a theory.

Whenever data or information is gathered, whether it be qualitative or quantitative, there are several broad questions that may be asked in relation to its use:

1. What data are needed?
2. How are the data to be used in various contexts?
3. How accurate are the data?
4. How should the data be expressed?
5. How are the data to be interpreted?
6. How are the data to be used in decision making?

Each of these shall be discussed in turn.

What data are needed?--Specification. The problem of the specification of measurement is the problem of deciding what is to be measured and under what circumstances. This is dependent on what are the concerns and questions of those who will be making the decisions based on the measurements. The nature of the decisions to be made will determine what measurement data will be collected. How extensive the collection of measurement data will be will also depend on the resources available such as money, time, and manpower.

An experimental design specifies the various treatment conditions under which measurements are to be made. An experimental design will also specify when measurements are to be made--pre-treatment, post-treatment or at regular stages during the treatment. A sampling design defines the population of objects or events whose property is to be measured. A sampling design also details how the sample on which measurements will be made is to be chosen from the population.

How are the data to be used in various contexts?--Generalizability. The question here is, if a certain measurement of a property is made under certain conditions, how will this measurement be affected by another set of conditions under which a decision has to be made?

The fact that we want to compare a measurement made under one set of circumstances with that made under another set, has led to the notion of standards. Churchman (1959:88) explains:

> The necessity for standards of measurement is based, in part, on an almost obvious observation that not all human experience takes place at the same time or in the same circumstance. Even if there were but one mind in all the world, such a castaway would need to compare the experience of one moment and place with that of another moment and place. He would have to communicate with his own past. The devices that men have used to make these comparisons are many indeed. One of the most direct methods consists of reconstructing each experience into an experience of a given moment and a given time, i.e., the

> present experience is "adjusted" into the experience that would have taken place under some standard set of conditions. This is the only way in which experiences of various moments can be communicated, but it is a very powerful device for communication. Robinson Crusoe cannot bring along his hut as he searches for a flagstone for his hearth. But he does need to compare an experience on the beach with a past experience in his hut. He does this (say) by the use of a piece of string. He argues that if the string length fits the flagstone, the flagstone will fit the hearth. What he is really saying is that each experience--of the hearth and the flagstone--can be adjusted to a comparison with the string under "standard" conditions.

The need for comparability from one situation to the next led to the notion of units of measurements. For example, a basic unit of length, widely used from earliest recorded history until the nineteenth century, was the cubit--the length of the forearm from the point of the elbow to the tip of the outstretched middle finger. This unit lacked high precision since it clearly varied with the size of the person involved. Fundamental units in the physical sciences include, besides the meter (length), the kilogram (mass), the second (time), and the ampere (electric current).

Why are the "measurements" given by tests not expressed in terms of units? The reason lies with the basic procedures for determining the empirical relationships between quantities of properties. The three basic procedures were described earlier in this chapter. The first or ordinal procedure is concerned with the arrangement of a property, p, in order according to size. That is, the relationships "greater in p than," "equal in p to," and "less in p than" are determinable independently of any measuring procedure. Educational and psychological tests aim (whether they are successful is another question) to achieve this first procedure. The second procedure is concerned with the counting of units using standard sequences and involves the concatenation of objects so

that the property of interest is combined (see discussion earlier in this chapter). It is not possible to carry out this concatenation process for the properties assessed by tests. If concatenation is not possible, then it is not possible to have units.

One aspect of how data are used in various contexts is generalizability. Generalizability is made possible if there are laws or known relationships between variables which allow one to adjust a measurement made under one set of conditions to another set of conditions. Laws typically exist in the natural sciences, whereas least squares estimates of the relationships between variables are typically used in the social sciences. The gas laws are prime examples from the natural sciences. For a given mass of gas, and as long as pressures are not too low, then

$$\frac{P_1V_1}{T_1} = \frac{P_2V_2}{T_2}$$

where P_1, V_1, and T_1 are the pressure, volume, and temperature on occasion one, and P_2, V_2, and T_2 are the pressure, volume, and temperature on occasion two. For example, if the temperature of a given mass of gas is known when the pressure and volume are P_1 and V_1, then the temperature can be calculated when the pressure and volume are P_2 and V_2. In the social sciences, regression equations are typically used to state the relationship between variables. For example, given a measure or measures of a student's achievement in the high school context, a regression equation is used to calculate an estimate of the student's achievement in the college context. Thus laws and regression equations allow one to adjust a measurement made under one set of conditions to another set of conditions.

Cronbach, Gleser, Nanda, and Rajaratnam (1972) discuss generalizability theory in relation to test scores. Generalizability theory examines how test scores vary under changing conditions. Generalizability theory represents a marriage between the factorial experimental design first developed by R. A. Fisher and classical test theory. In a generalizability study, one obtains two or more scores for a person by observing him under different conditions, and examines the consistency of the scores. The analysis estimates components of variance, each attributable to one condition or combination of conditions represented in the experimental design. These estimates may show, for example, that one form of a test elicits about the same behavior as another parallel form of the test, but that variations in test behavior from one testing to the next are substantial. Generalizability theory is conceived in terms of the accuracy of tests under various conditions. Consequently generalizability theory is also discussed in the next section concerned with the accuracy of measurement.

How accurate are the data? "Accuracy is itself a measurement--the measurement of the degree to which a given measurement may deviate from the truth. No procedure can claim the name of measurement unless it includes methods of estimating accuracy" (Churchman, 1959:92).

Accuracy is a highly relative term, from at least two perspectives. One can look at accuracy in terms of the percentage error of measurement. An error of six centimetres in measuring the width of a desk 60 centimetres wide is a large error, being ten percent of the total. However, the same error of six centimetres in measuring the distance between the earth and the moon is infinitesimal. A second way of looking at accuracy is in terms of the decisions or actions that have to be made on the basis of the measurement. An error of six centimetres

in the width of a desk can be important in deciding whether the desk can be taken through a doorway, whereas an error of one millimeter is inconsequential. However, this is a simple example where it is relatively easy to decide how serious a certain size error will be for the decision that has to be made. Often it is difficult to decide how serious the size of the error is, especially if one is not very clear about what decisions or actions have to be made on the basis of the measurement. The problem of accuracy is made more difficult if a measurement is to be used in different contexts. Churchman (1959:92) states:

> In statistical literature, accuracy is sometimes defined in terms of a "confidence interval." In so far as this computed interval has any meaning, it tells us that a certain range of numbers constructed out of observations has a specific probablity of including the "true" measurement. Each set of observations is the basis for forming a net to "catch" the truth, and the confidence interval tells us the probability of a successful catch. But it is almost always difficult to determine how the information supposedly contained in a confidence interval can be used; i.e., what difference would it make if the confidence interval were twice as large, or half as large? Most statisticians seem to prefer to negotiate this tricky question by urging the decision maker to set his own size of confidence interval. Since most decision makers honestly do not see the purpose of the interval in the first place, the interval is set "arbitrarily," i.e., pointlessly.

The decision problem of accuracy has not been solved. The problem of accuracy is to develop measures that enable the measurement user to evaluate the information contained in the measurements.

One approach to achieving accuracy of measurement is to carry out measurements by standard procedures under standard conditions. For example, for measuring the amount of a certain chemical present in a sample, standard procedures of chemical analysis are laid down, especially when such chemical measurements are to be reported

"officially," such as in a court of law. Educational and psychological tests are administered under standard procedures and conditions such as time limit, oral instructions to subjects, preliminary demonstrations, ways of handling queries from subjects, and every other detail of the testing situation. This standardization is aimed at controlling the conditions that could affect the measurement so that these conditions are the same from one measurement to the next. This leads to the notion in test theory of the reliability coefficient which is an indicator of accuracy.

This section and the previous one look at many different aspects of the measurement process. But of all these aspects, the one aspect that psychometricians have chosen to focus most of their energies on is accuracy. Cronbach et al. (1972:23) state that "the heart of traditional measurement theory is the so-called reliability coefficient, the ratio of 'true score' variance to observed-score variance." As Lumsden (1976:251) indicates, the most highly regarded notion in all test theory, and the only one to be seriously developed, has been the venerable, observed score equals true score plus error. The major purpose for this decomposition is to provide a rationale for the reliability coefficient.

There are several different types of relaibility coefficients corresponding to different types of possible error. One method for finding the reliability of test scores is by repeating the identical test on a second occasion. The reliability coefficient in this case is simply the correlation between the scores obtained by the same persons on the two administrations of the test. Retest reliability shows the extent to which scores on a test can be generalized over different occasions. The higher the reliability, the less susceptible the scores are to the random

daily changes in the condition of the subject or of the testing environment.

Another reliability coefficient is the parallel-form type. The same persons can be tested with one form on the first occasion and with a parallel form on a second occasion. The correlation between the two sets of scores represents the reliability coefficient. If the two forms are given immediately following one another, the major error will be due to content sampling. "Everyone has probably had the experience of taking a course examination in which he felt he had a 'lucky break' because many of the items covered the very topics he happened to have studied most carefully. On another occasion, he may have had the opposite experience, finding an unusually large number of items on areas he had failed to review. This familiar situation illustrates error variance resulting from content sampling" (Anastasi, 1976:113). If the parallel forms are not given immediately following one another, then the reliability coefficient indicates not only error due to content sampling but also error due to sampling over occasions as for test-retest reliability.

Split-half reliability is obtained from the single administration of a test. The test is split into halves so that the two resulting forms are comparable. Two scores are thus obtained for each person and the correlation is found between the resulting two sets of scores. The split-half reliability coefficient (corrected by the Spearman-Brown formula) is an indicator of error due to content sampling. A similar type of reliability coefficient is the Kuder-Richardson reliability coefficient, there being two versions, Formula 20 and Formula 21. A test can be split into half in a large number of ways, and for each of these splits, the split-half reliability coefficient can be calculated. The mean of all such split-half coefficients would give the Kuder-Richardson

reliability coefficient. The Kuder-Richardson reliability coefficient applies only to tests whose items are scored "right" or "wrong" and a more general coefficient has been developed, known as coefficient alpha.

How do reliability coefficients as indicators of accuracy aid interpretations and decision making? First, the reliability coefficient is one of the criteria used in guiding decisions about test selection. Second, the reliability coefficient can be used to make regression estimates of true scores and the standard error of measurement.* Third, the reliability coefficient can be used to correct a validity coefficient such as the correlation between a test and the criterion it is designed to predict. This is known as the correction for attenuation.** Lord and Novick (1968:71) state:

> The idea is that the correlation between observed scores is less than the correlation between corresponding true scores because the former correlation is attenuated by the unreliability of the measurements. If the reliabilities of the measurements are known, then . . . formulas may be used to compute the disattenuated correlations, i.e., the correlations between the corresponding true scores. Attenuation theory is one important justification for the emphasis that classical theory has placed on the concept of reliability.

Generalizability theory, proposed by Cronbach, Gleser, Nanda, and Rajaratnam (1972), represents the culmination of a decade of work on reliability seen as generalizability. Generalizability theory represents a marriage between the notions of accuracy and generalizability. A

**Lumsden (1976:256) states, "The correction should never be used. It too often produces corrected correlations which are greater than one, and it is not sufficient to pass these occasions off with an embarrassed smile and some mutterings about unreliability of estimates."

*Linear regression estimates of true scores may be quite misleading if regression is not linear. Setting up confidence limits using the standard error of measurement will also be misleading since the standard deviation of error scores is not independent of true scores (see Lumsden, 1976:255).

behavioral measurement is seen as a sample measurement from the collection of measurements that might have been made. One could take an average of the collection of measurements and this is terms the universe score. The difference between the observed score and the universe score is taken to be error. The universe score is taken to be analogous to "the true score" of classical test theory.

There are many different universes one might generalize to. Any person fits within many different populations. John Smith may be considered as a sample from any of several sets: residents of Illinois, plumbers, persons with a $30,000 income, Democrats, etc. Any measurement likewise fits with a variety of universes of conditions. Any measurement is carried out under a set of conditions: the time of day, in a particular physical setting, with a particular observer, with a particular set of stimuli, etc. The general term referring to conditions of a certain kind if facet. Thus, observations may be classified with respect to the facet of days of testing, the facet of settings, the facet of observers, etc. A universe of observations will be characterized with respect to one, two, or more facets.

Why would one want to generalize over a set of conditions? Would one not be more interested in a measurement made under a particular set of conditions? Why would one be interested in an average of measurements made over sets of conditions? Why would the observed score under a particular set of conditions be seen as in error from an average score taken over a universe of conditions? Cronbach et al (1972:21) claim there is a need for a universe score and cite an example:

> The universe to which an observation generalizes depends on the practical or theoretical concern of the decision maker. Consider a supervisor's rating of an employee. This rating differs from what would be recorded on another occasion, since the supervisor's mood at the time of

> rating and his recent experience with the employee have some transient effect. The investigator concerned with employee effectiveness surely wants to generalize over the class of ratings the supervisor might have given at other moments. The investigator will generalize over a time period of perhaps a month if the rating is taken as an end-of-year report of the employee's qualities. Any of the moments within that month would presumably have been a suitable time for the inquiry. In another study, where the rating is a datum for an intensive study of week-to-week changes in supervisor attitudes during a human-relations course, the investigator will generalize over only a single day. If the rating is a criterion against which he will validate an ability test, he needs to generalize over supervisors as well as occasions. But if the sole concern is whether the employee is getting along with <u>this</u> supervisor, the universe of possible supervisors is irrelevant.

Within the natural sciences, there appears to be no corresponding need for generalizability theory. In the natural sciences, the interest is in an average of measurements made under a <u>particular</u> set of conditions rather than in an average of measurements made over <u>sets</u> of conditions. For example, in the practical science of engineering, which is highly decision oriented, there is an interest in particular conditions. What is the greatest weight this bridge will have to bear? What is the highest wind this building will have to withstand? What are the extremes of temperature this telephone wire will exist under? In engineering, a measurement under a <u>particular</u> set of conditions is not viewed as in error from an average of measurements taken over a universe of conditions. In fact, taking such a view in engineering would be ridiculous. Cronbach et al. (1972:21) justify their view with the example of a supervisor rating an employee over sets of conditions. However it seems just as likely that there would be an interest in the employee's performance under <u>particular</u> sets of conditions. It is difficult to understand why the social sciences need generalizability theory.

How should the data be expressed?--The language of communication.

Churchman (1959:85) states:

> The measurer must develop a language which adequately communicates to another person what the user must do to utilize the information contained in the measurement. The emphasis here is on the language of communication
>
> One aim of the language of measurement is to communicate to as many potential users as possible since this will increase the scope of utilization. Another aim is to enable the user to employ the information when there is need for fine distinctions since this also will increase the scope of utilization. These two aims are apparently in conflict--the more common the language, the more difficult it is to use the language for portraying fine distinctions.

In the case of achievement tests, for example, various aids have been developed to help the test user interpret test scores. By itself a test score means very little. For norm-referenced tests, tables of norms are supplied to give meaning to a test score and to aid the test user in decision making. For a criterion-referenced test, a criterion or mastery level score is given to which a test score can be compared. For a domain-referenced test, the test score of a student can be interpreted in terms of the percentage of the domain that has been achieved by the student. However, communication of a test score in terms of some form of norms is most common. Examples are grade equivalents, percentiles, and standard scores (e.g., z scores, T scores, stanines).

While a test score has a precision about it, it also represents a loss of information since a test score involves data reduction. A test score does not specify exactly what a student can do. A test score results from summing the scores on a number of items. A test score does not inform the test user about the student's performance on individual items.

How are the data to be interpreted? As indicated in the previous section, the way in which a measurement is communicated will determine how it is interpreted. But interpretation of the measurement of a variable is also dependent on how the measured variable is related to other variables. For example, the outside temperature can be interpreted in terms of what amount of clothes to wear in going outside. The weight of an object is an indicator of how difficult it will be to move the object from one place to another. The meaning and interpretation of a variable is dependent on what other variables it is related to. The relationship between variables is often expressed in terms of laws or theoretical networks.

Test validity is the area of test theory that deals with the interpretation of measurements. Cronbach (1971:447) states:

> Validation examines the soundness of all the interpretations of a test--descriptive and explanatory interpretations as well as situation-bound predictions.
>
> To explain a test score, one must bring to bear some sort of theory about the causes of the test performance and about its implications. Validation of test interpretations is similar, therefore, to the evaluation of any scientific theory. . . .
>
> One validates, not a test, but an interpretation of data arising from a specified procedure. A single instrument is used in many different ways--Smith's reading test may be used to screen applicants for professional training, to plan remedial instruction in reading, to measure the effectiveness of an instructional program, etc. Since each application is based on a different interpretation, the evidence that justifies one application may have little relevance to the next. Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test "is valid."

Thus test validation is viewed as the validation of interpretations of data arising from various test administrations.

How are the data to be used in decision making? Cronbach (1976:200) views "improved decisions as the aim of measurement." Churchman (1959:84) states:

> Measurement is a decision making activity, and, as such, is to be evaluated by decision making criteria.
>
> In this sense, i.e., measurement taken as a decision making activity designed to accomplish an objective, we have as yet no theory of measurement. We do not even know why we measure at all. It is costly to obtain measurements. Is the effort worth the cost?

However, in relation to tests, Cronbach and Gleser (1965) have put forward a theory of measurement as a decision making activity. The theme of their book is that the practical uses to which a test is put and the measurement procedure should not be separated. When decisions are based on test scores, it is the accuracy of classification of a person which matters so that precision of measurement is valuable only insofar as it enhances the quality of decision, i.e., reduces misclassification. Two special cases of classification are important in relation to tests--placement and selection. An example of a placement is dividing students among sections to be taught at different rates and using a test for coarse grouping of applicants. If the decision is acceptance or rejection into a treatment, then this case is one of selection.

Cronbach and Gleser (1965) stress that information must be purchased and that costs should be assessed as accurately as circumstances permit. Cronbach and Gleser discuss the utility of information as a function of the benefit of using the information as the basis for decisions minus the cost of gathering the information. They also emphasize the bandwidth-fidelity dilemma. The test designer and the user of tests frequently have to choose between careful estimation of a single variable and more cursory exploration of many separate variables. In any decision

situation, there is some ideal compromise between variety of information (bandwidth) and thoroughness of testing to obtain more certain information (fidelity). For example, if a person is choosing published tests for a testing program, he has to decide whether to use the available time to measure one or two variables by means of long tests, or employ a much larger number of short tests measuring a variety of characteristics. In 1976, Cronbach (1976:200) reviewed the value of decision theory (D theory).

> We have had D theory for two decades; in that time, its formulas have almost never been applied to actual data. To apply D theory formally demands data that can rarely be assembled. At best, the machinery yeilds precise analyses of test efficiency that have little practical advantage over rough estimates. The important aspect of D theory has been the questions it put into our minds. I offer several examples.
>
> 1. Formerly, measurers equated adequacy of measurement with test length. D theory advised us to consider the cost-effectiveness of the design for a measuring instrument. It advised us to reduce our concern for precision as such, and to ask, instead: To what degree if any does collecting additional information improve a decision?
>
> 2. When we began to see improved decisions as the aim of measurement, that insight changed our view of other matters. An evaluation, we now say, is a study in aid of decisions. . . .
>
> 3. D theory led us to see that most decisions about students are placement decisions, and that these can be validated only by the study of Aptitude X Treatment interaction. . . .
>
> 4. D theory's concept of utility of information continues to prove its value. It has recently transformed the discussion, among measurement specialists, about bias in selecting students and employees.

To summarize this section, the first column of Table 1 shows the various aspects of the measurement process. Section A refers to the four steps developed in the previous section, and Section B shows the six aspects of using measurements. In the second column, the corresponding

aspects of test theory are listed. Opposite A3 and A4, the aspects of test theory have been placed in parentheses, since these aspects of test theory do not fully reflect A3 and A4. It is clear from Table 1 (p. 42) that test theory has largely developed around the use of measurements. There is nothing to correspond to A1 and A2. The conclusion is that, if tests are to result in measurement, there is need for a development in test theory around the four aspects of obtaining measurements which are associated with the nature of tests. This completes the discussion of measurement. Other ways of assigning numbers will now be discussed.

Counting

If one had to choose the way numbers were historically first assigned to objects and events, one would choose counting. Enumeration, that is counting, is a very basic process and is something that we commonly carry out in our everyday lives. It seems so simple that it is briefly and rarely discussed in the literature as a method of quantification.

Numerosity, number and counting, though they are linked, can be distinguished from one another. Numerosity, the property of a group of entities or attributes, when it is conceived as "many" or "few," is a classificatory concept and is conceived without the use of number or counting. This is also true when numerosity is a comparative concept, as when it is conceived as "greater than, less than, or equal in manyness." Nelson and Bartley (1961:179) argue that numerosity can be discerned without the use of number or counting:

TABLE 1

A SUMMARY OF THE VARIOUS ASPECTS OF THE MEASUREMENT PROCESS SHOWING THE CPRRESPONDENCE WITH TEST THEORY

Aspects of the Measurement Process	Aspects of Test Theory That Purport To Correspond
A. Obtaining Measurements	
1. Understanding the property in qualitative terms.	
2. Conceptualizing the property in quantitative terms.	
3. Development of a procedure for determining the empirical relations between amounts of the property.	(Test construction)
4. Assignment of numbers so that resulting numerical relational structure is homomorphic with empirical relational structure.	(Test scoring) (Scaling theory)
B. Using Measurements	
1. What measurements are needed?	Decision Theory (Cronbach and Gleser, 1965)
2. How are the measurements to be used in various contexts?	Generalizability Theory (Cronbach et al., 1972)
3. How accurate are the measurements?	Reliability Classical Test Theory (Lord and Novick, 1968) Generalizability Theory (Cronbach et al., 1972)
4. How should the measurements be expressed?	Norm, criterion, and domain referencing
5. How are the data to be interpreted?	Test validation
6. How are the data to be used in decision making?	Decision Theory (Cronbach and Gleser, 1965).

> One might first ask whether there are conditions under which discrimination of "number" can be made without knowledge of systems of thought involving number. There are. It is known that youngsters are capable of perceiving the "manyness" difference between 5 oranges and 20 marbles before formal indoctrination in arithmetic. This testifies to the fact that there exist "natural classes" of one, two, three, etc. objects, which, if they are not too large, are something perceivable or directly discriminable. Taves (1941) used exposure times too short to permit counting and found adults able to compare the "manyness" of dots accurately providing they did not exceed ten in number. In addition, such natural classes are items that are perceivable even by various sub-human species. O. Koehler (1956) has shown this to be the case in pigeons in which the experimental variable is actual number of items, other stimulus factors varying at random. Under these conditions his pigeons did discriminate, and this discrimination had to be based upon such natural classes. Let us say then that these natural classes have numerosity. It is numerosity, not number, that is discriminative in nature.

Numerosity, conceived as a comparative concept, can also be discerned without the use of number or counting. Consider two boxes, each containing prepared biological slides. It is possible to tell without number or counting whether the two boxes "have the same numerosity" of slides, or whether one "has more than" and one "has less than." The procedure would be to remove one slide from one box and at the same time remove one from the other box. This double operation could be repeated until one, or both, of the boxes were emptied of slides. If both boxes are emptied of slides, then the boxes had the same numerosity originally. However, if one of the boxes still contains slides after the other has been emptied, then that box had a greater numerosity of slides to begin with.

Perception and comparison are not the only ways that numerosity may be discriminated. When numbers are associated with numerosity, numerosity is being conceived as an absolute concept. Numerosity and number can be distinguished. Numerosity is a natural fact which has to

do with perceived or inferred manyness. On the other hand, numbers, while they may represent numerosity, differ. Numbers have an independent existence, as, say, marks on paper, from events with numerosity. They can be treated symbolically as though they are completely divorced from natural numerosities. While number and numerosity differ, number is used interchangeably with numerosity in everyday language. For example, Ellis (1966:152) states, "The word 'number' is certainly used as a quantity name. We talk of groups being equal in number, or of one group being greater in number than another, just as we talk of objects being equal in length, or of one object being longer than another."

Counting is the process by which the number indicating numerosity is arrived at. Counting provides an alternative to sensory discrimination when the task is one of detecting numerosity. Nelson and Bartley (1961:181) state:

> In counting, the class of natural numbers beginning with one and ordered in the usual way is used to interpret a class of objects. Members of the class of objects taken one by one in any order are put, one by one, in correspondence with the ordering of numbers. The last number of course names the class of objects. Counting, insofar as it is a numbering device, is arbitrary in the sense that it involves human convention.

Counting requires that a group of entities or attributes be formed and that there must be at least implicit criteria why members belong to a group. Members of a group need not be identical but they must be equivalent in some way. Smith (1938:4) gives a clear example of what is meant by equivalence. He also explains how counting (i.e., enumeration) is different from measurement:

> Suppose we have a pile of boards and we wish to know how many boards there are in the pile. We enumerate them and discover that there are thirty-two. Enumeration requires that some class of objects be defined. The definition does not require that the object have identical properties nor that they possess the properties in the same degree.

> The boards in the pile may vary in color, temperature, length, weight, density, hardness, and so on, but they must be objects near enough to constitute a single class of objects which we call boards. Thus by enumeration we can answer the question: "How many boards?" If we ask, however, whether one board is longer or heavier than another, obviously we have raised a problem that enumeration cannot answer. To answer this question requires measurement, because enumeration cannot be used to describe the different degrees of property such, for example, as lengths and weights of boards. We can enumerate the boards, but we cannot enumerate their lengths and weights. It is precisely the variation of the properties which, ignored by enumeration, constitutes the area dealt with by measurement. Measurement is not concerned with a class of objects nor with the question of how many objects of a particular class are present at a given time and place.

Ellis (1966:153) agrees that counting is not the same as measurement:

> It is clear that we should say that number is a quantity. But on the other hand, it seems to make no sense to speak of measuring the number of things in a group. We can speak of counting, calculating, or guessing the number, perhaps, but not of measuring it. . . . The verb "to count" can be used transitively or intransitively. Intransitive counting, that is the idle recitation of the numeral sequence, is admittedly not very like measuring. If we are measuring, then we must always be measuring something in some respect. We cannot simply be measuring. But transitive counting has some of the essential characteristics of a measuring procedure. For it is an objective and determinative procedure for assigning numerals to groups--("objective" in the sense that any one who follows the same procedure with sufficient care will be led to assign the same numerals to the same groups).

In an earlier example, two boxes of biological slides were compared to one another. This procedure would have worked just as well were the objects involved of a different kind. The biological slides could have been compared to the burners on the laboratory bench or the number of bottles of chemicals on the shelves. This leads to the following two rules that are so obvious that they are seldom noticed.

1. If two sets of objects, when compared against a third set, are found to have the same number as the third set, then, when counted

against each other, they likewise, will be found to have the same number. This rule enables us to determine whether two sets of objects have the same number without actually bringing them together at all. It suggests the possibility of portable standard collections which can be counted, first against one collection then against another, in order to tell us whether these have the same number (Hanson, 1969:45).

2. Hanson (1969:45) suggests another rule, ancillary to the first: By starting with a single object and continually adding another object to it, we can build a series of collections, e.g., ., .., ..., ..., ..., of which some one collection will have the same number as any other collection whatever. This rule suggests the efficacy of counting collections not against each other, but against some single standard collection.

Counting differs from the measurement of both interval and ratio concepts in that there is nothing corresponding to a standard unit. The numerical sequence itself plays a role similar to that of a standard. In the process of counting, members of a group of entities are matched with members of the numerical sequence. Thus, if a group of five entities is counted, the subset of the numerical sequence, 1, 2, 3, 4, 5, serves as a standard group of five. Ellis (1966:156) expresses it this way:

> It is conceivable that we would use groups of stones or marbles as numerical standards, and that they should be kept in special museums to protect them from destruction. All number determination would then be done by matching the group whose number is to be determined with one of the standards, or its equivalent in number. We can imagine, indeed, that we should all carry around little bags of marbles labelled 2, 3, 4, and so on, and that when we wish to find the number of things in a given group, we simply find the bag which contains the same number of marbles. This would be logically very similar to carrying around a set of feeler gauges. In fact, of course, this would be very cumbersome, and it is much easier to use subsets of

the numberical sequence as numerical standards. The subset of numerals 1 to 12 is a far more convenient numerical standard than a bag of 12 marbles.

Primitive man probably used his fingers and toes as numerical standards. According to Hanson (1969:45), less primitive man used standard collections of "counters," small bead-like tokens, of which a great many could be carried at once in a single bag. The beads of an abacus serve as numerical standards. Parenthetically, it can be noted that in shops we still say that we conduct business "over the counter." Numerals serve the purpose of our counting tokens. Numerals are just the distinguishable "objects" out of which we build our standard series--by adding them in turn to previous members of the series: "1," "1, 2," "1, 2, 3," "1, 2, 3, 4," and so on, indefinitely. We count other collections against these members of the standard series. In this way, we can ascertain whether or not two collections so counted have the same number.

In the measurement of interval or ratio concepts, the choice of unit is arbitrary. For example, to form a scale of length, all that is necessary to do is to select any object which bears a stable length relationship to sufficiently many other objects. This object can serve as the unit of length and be assigned a numeral of one. All other lengths will be some fraction or multiple of this length. The unit must be given a name. It is meaningless to say, "This object has a length of five." It is necessary to say that it is 5 meters or 5 feet, indicating the name of the standard of length that has been chosen.

In contrast with enumeration, there are no such choices to be made. There is no arbitrary selection of a unit to act as standard. If oranges are being counted, there is no arbitrary selection of a certain group of oranges to act as a standard. Nor is the invention of a unit meaningful. It is not meaningful or useful to say about a group of five

oranges that, "This group contains five gronks," where "gronks" is an invented name. In order to carry out the process of counting, it is necessary that the entities form a group because of their equivalence in some respect. Because of their equivalence, any one of the entities in a sense is a unit.

The number given by a measurement procedure is dependent upon the scale of measurement used. Whether one counts by ones, twos, or threes, the number of entities in a group is quite independent of the counting procedure used. The statement, "This group contains five entities," has a precise meaning which is independent of counting procedures.

Counting is the process of attaching a number to the numerosity of a group of entities or attributes. Counting requires a classificatory concept since one must be able to decide whether the entities or attributes do or do not belong to the group. Thus a very important question to be asked is, "Are the entities or attributes equivalent or are they different?" The question appears quite trivial when apples or oranges are being counted, but the question is quite a difficult one when phenomena about which little is known are being studied. An example will be taken from the history of science as illustration.

Astronomy was one of the earliest sciences. The heavens were rich in phenomena. To the ancients the stars, planets, comets, meteors, thunder, lightning, hail, clouds, and rain were more alike in their origin and effects than they are to us today. They found what was in the sky as inaccessible as we find aspects of the human mind today. The word "meteorology" (the study of the-things-on-high) covered at first the science of all those things which happened above the earth and so were inaccessible to close inspection (Toulmin and Goodfield, 1961:17). Today we restrict the same word to the science of the weather, knowing that

climatic happenings have to be explained differently from astronomical happenings. Thus part of saying things are different is explaining them in different terms.

Simple observation is not enough to say whether entities are the same or different. Astronomy was rich in observation for thousands of years. Theories and systematic investigation are needed to explain whether phenomena belong to the same category or not. To the ancients it was not clear that planets were essentially different objects from the stars even though it was observed that the planets moved in relation to the fixed stars. The sun appears so different from the stars, it was no wonder that it took so long for it to be realized that the sun and the stars belonged to the same nuclear phenomenon.

Thus the question of whether entities belong to the same category or different categories of phenomena is not necessarily an easy question to answer. We count up the items answered correctly on a test as if each item required essentially the same type of mental process. We have no theories to tell us if it is essentially like regarding planets and stars as belonging to the same phenomenon.

One limitation on counting is that it applies only to the discrete (Hanson, 1969:46). We cannot count the drops in a beaker of alcohol unless we first introduce a convention as to what will be regarded as "a drop." This agreed, we must then find a way of separating out these drops from one another in an invariant and uniform way. Similarly we cannot count electromagnetic radiation though we may be able to obtain a count on the number of wavelengths that pass a given point in a second.

It is common for the number obtained by an enumeration procedure to be used as an indicant of some property. A common example is that the number of items correct on a test is used as an indicant of an

individual's ability to correctly answer the items. That is, the number obtained by enumeration is regarded as a "measurement" of a property on some scale. When a number obtained by enumeration is used as an indicant of a property, it is still necessary for the enumeration process that the entities counted still belong to the one class. That is, the entities must still be equivalent in some way.

In summary, counting is the process by which the number indicating numerosity is arrived at. Counting requires that a group of entities or attributes be formed and that there must be at least implicit criteria why members belong to a group. Members of a group need not be identical, but they must be equivalent in some way. Counting differs from the measurement of both interval and ratio concepts in that there is nothing corresponding to a standard unit. In order to be counted, entities must belong to the one group. It is not always easy to decide whether entities belong to a group or do not. We count up the items answered correctly on a test as if each item required essentially the same type of mental process. We have no theories to tell us if it is essentially like regarding planets and stars as belonging to the same astronomical phenomenon. The number obtained by enumeration is often used as an indicant of some property, as is the case with tests.

Naming By Numbers

Naming by numbers does not constitute measurement or counting. Examples of naming by numbers are telephone numbers, numbers on football jerseys, or social security numbers. Naming by numbers is sometimes called nominal measurement, but it is not intended that the numbers indicate empirical relations as for measurement. For example, a football

player with the number 40 has not twice the ability of a player with the number 20. Nor do the numbers have anything to do with counting, as for example how many members are on the football team. Furthermore, the numbers do not indicate any order or rank, such as the order of their ability at football or the order in which they will play.

A set of words may have the same denotation as the numbers. For example, a particular galaxy is identified as "M31" or "NGC224," according to its listing by Messier or in the New General Catalogue. It is also known as "The Great Nebula in Andromeda."

A number may be used as a label and at the same time indicate some kind of rank. Thus four courses in chemistry might be given the labels Chemistry 100, Chemistry 200, Chemistry 300, and Chemistry 400. At the same time the numbers are used to indicate rank in terms of level of advancement in chemistry.

When numbers are assigned as social security numbers, to licenses, or to football players, no two numbers are assigned to the same object, and no two objects have the same number assigned to them. This does not hold when numbers are used as labels to categories. For example, in a swimming class, the students were regularly divided into two groups, Group One and Group Two. Group One swam across the pool first and Group Two swam second. In this case two persons could have the same number assigned to them. In addition, the labels also indicated the rank or order in which the groups swam across the pool.

When classes are given a number label, the number will also indicate some order according to whether the classes are ordered or not. For example, for the purpose of statistical calculation, males are often given the label One and females the label Zero. No order is intended in these labels since there is no order in the categories, male and female.

One way of regarding a rating scale is as a set of ordered categories. A very common form of rating scale requires the rater to judge which of five classes his agreement-disagreement falls into, as the following example illustrates.

The instructor was very helpful to me

5	4	3	2	1
Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree

The numbers can be regarded in two different ways. First, the numbers could be regarded as indicating the classes of agreement-disagreement. When regarded in this way a frequency distribution showing the number of people falling into each class could be given. Second, agreement-disagreement could be regarded as a comparative concept, the higher the number, the higher the amount of agreement. In this case, a median or mean is commonly calculated. So with a rating scale which uses numbers, the numbers may be regarded as classificatory labels or the numbers may be regarded as indicating a comparative concept involving human judgment of order.

Numerical Judgment

The methods of numerical judgment are very well covered by Torgerson's book (1958), Theory and Methods of Scaling. They will not be dealt with in detail here, but a brief summary of the major methods will be provided.

Toregerson (1958:61) describes one set of methods as the "subjective estimate methods."

> The rationale is basically the same regardless of the subject-matter area in which the methods are used. The procedures are as follows: a series of stimuli is presented to a subject (judge, observer), who is

> instructed to render a direct, quantitative judgment of the amount of a specified attribute that is possessed by each of the stimuli. For example, the task set for the subject might be to rate each of a group of essays on a ten-point equal-interval scale with respect to excellence. It is assumed that the subject is capable of carrying out this task. That is, if he rates three essays as 4, 5, and 6, respectively, it is assumed that, except for a certain amount of error, the difference in the degree of excellence between the first and second essay is equal to the difference between the second and third. He is thus required to make his reports so that they may be treated as scale values on a linear scale of the attribute of interests.

Numerical judgment is associated with comparative concepts. That is, in contrast to Torgerson, who assumes judgments can be made on an interval scale, a conservative view assumes that numerical judgments can be made only on an ordinal (comparative) scale. Judgments can be made of unidimensional concepts or multidimensional concepts. As an example of a unidimensional concept, consider the following experiment: The subjects are presented with a series of objects differing only in weight. They are instructed to rate these objects on a ten-point, subjective scale of weight, with the higher digits corresponding to the heavy end of the continuum, and the lower digits to the lighter end. The subjective concept, weight, is unidimensional. However, grading essays involves making a subjective estimate of a multidimensional concept, since many different aspects--expression, grammar, punctuation, content, etc.--are taken into account.

Numerical judgements can also be obtained by fractionation methods which assume that a subject is capable of directly perceiving and reporting the ratio of two subjective magnitudes. Fractionation methods exist in two different forms. With direct-estimate methods, the subject is presented with two stimuli and instructed to report the subjective ratio between them with respect to the designated attribute. for

example, two tones of the same pitch might be presented to the subject with instructions to report the ratio of loudness of the first tone to the second. With prescribed-ratio methods, one stimulus (the standard) is kept fixed, and the other (the variable) is adjusted. The subject's task is to report when the subjective ratio of the variable to the standard is equal to some prescribed ratio. For example, the subject might be instructed to adjust the variable stimulus until it is one half as loud as the standard. Fractionation methods have been carried out with the unidimensional concepts--loudness, pitch, numerosity, weight, saltiness, sweetness, sourness, bitterness, brightness, and time intervals. If it can be assumed that the subject is capable of carrying out the instructions, then a ratio scale is obtained. Torgerson (1958:113) comments, "As was true of the subjective-estimate methods, we have no basis for concluding that the scale actually possesses the properties attributed to it from the data gathered in the scaling process itself. Scale values of the stimuli could always be computed, assuming only that the subject could make valid ordinal judgments and that he behaved more or less consistently."

With equisection methods, the subject has available an unlimited number of stimuli which are ordered with respect to the attribute. The experimenter designates the smallest and the largest as standards. The task set for the subject is to select n-1 of the remaining stimuli so that the n-1 stimuli divide the interval between the two standards into n subjectively equal intervals.

The differential-sensitivity methods are based on the concept of the equality on the psychological continuum of just noticeable differences. That is to say: If a stimulus B is "just noticeably greater" than stimulus A, and stimulus C is "just noticeably greater" than stimulus B,

then the distance on the psychological continuum that separates A and B is equal to the distance separating B and C. This notion may be taken as an assumption, subject to further test, or as a definition of what is meant by equality of intervals on the psychological continuum.

In the method of paired comparison, each stimulus is paired with each other stimulus. Each pair is presented to the subject, whose task is to indicate which member of the pair appears greater (heavier, brighter, louder) with respect to the attribute to be scaled. The subject must designate one of the pair as greater. No equality judgments are allowed. It is possible mathematically to calculate the scale value of each of the stimuli.

There are at least three sets of procedures where stimuli are placed into categories which are ordered with respect to the attribute being investigated. The first set of procedures is sorting procedures, which includes the method of successive intervals. The subject's task is to sort the stimuli into piles so that the first pile contains those stimuli that are most positive with respect to the attribute; the second pile, the stimuli next most positive, etc. It is only necessary that the piles be in rank order with respect to the attribute. Often the piles may be identified with adjectives which progress from extremely positive to zero or extremely negative, depending on the particular attribute. The second set of procedures is rating procedures. The stimuli are presented one at a time, instead of presenting all stimuli to the subject at once. The rating may be expressed on a numerical scale (e.g., rate on a scale from 1 to 5, 5 being most positive) or an adjective scale. The numerical scale (not the actual rating) is an example of naming ordered categories by numbers. The third set of procedures is ranking procedures. The subject is required to place the stimuli in rank order with respect to

the attribute. Each rank may be taken as a category, or several adjacent ranks may be combined to make up each category.

This completes the brief description of the various methods of making numerical judgments as described by Torgerson (1958). It is conservatively assumed that numerical judgment can be made only on an ordinal scale, i.e., numerical judgment is associated with comparative concepts and not with interval or ratio concepts. In general, the various methods described can be used for making numerical judgments of unidimensional as well as multidimensional concepts. It should be clear that numerical judgment is quite distinct from the other ways of assigning numbers previously described--measurement, counting, and naming by numbers.

Indices

At the outset, it is important to distinguish between an "indicator" and an "index." For example, "the number of times a person attended church in a year" is an indicator of his religiosity. An indicator will be taken as a unidimensional concept. On the other hand, if numbers are assigned to religiosity by making an algebraic composite of church attendance, number of other church-related activities, proportion of income spent on religious matters, and frequency of Bible reading, then this is called an index. Following Lazarsfeld and Rosenberg (1955:16) and Bojean, Hill, and McLemore (1967:2), an index is considered as an algebraic composite of several indicators. Hence an indicator is taken to be a unidimensional concept and an index to be a multidimensional concept.

A feature of both an indicator and an index is their indirectness in giving information about a characteristic. Thus the percentages of illiterates in a series of demographic units may be used as negative indicators of the general cultural level of the units which cannot be ascertained directly. The number of rooms in the dwelling of a family may be used as an indicator of the family's economic status. Occupational status, father's level of education, and number of books in the home can be combined to form an index of socioeconomic status of a family. In an intelligence test, each item correctly answered is taken as an indicator of intelligence. The number of correct answers on the test may be taken as an index of a person's intelligence. In physics, indicators are common. For example, temperature is not measured directly, but is indicated indirectly by the length of a mercury column, the resistance of a platinum wire, or the pressure of a constant volume of gas. Economists typically use indices to characterize general market-price movements. Perhaps the best known is the Dow-Jones Industrial Average, which is the total market price of one share each of 30 representative stocks, divided by 1.661 (Christy and Clendenin, 1974:225). According to Hagood and Price (1952:138), the direct-indirect division is not an absolute dichotomy and there are borderline cases which would be difficult to classify.

According to Lazarsfeld (1958:101), the process by which concepts are translated into empirical indices has four steps: an initial imagery of the concept, the specification of dimensions, the selection of observable indicators, and the combination of indicators into indices.

1. Imagery. Lazarsfeld (1958:101) describes imagery as follows:

> Out of the analyst's immersion in all the detail of a theoretical problem, he creates a rather vague image or construct. The creative act may begin with the perception of many disparate phenomena as having some underlying characteristic in common. Or the investigator may have observed certain regularities and is trying to account for them. In any case, the concept, when first created, is a somewhat vaguely conceived entity that makes the observed relations meaningful.

For example, the beginning idea of intelligence was observation and involvement with children--some strike one as being alert and interesting and others as dull and uninteresting. This kind of general impression starts the wheels of conceptualization rolling. Observations of social relationships might begin, for example, with reflections on one's own personal experiences in relationships and,, on this basis, specifications of the concept follow.

Beginning with imagery of a concept has linguistic problems associated with it. Payne (1975:35) suggests that imagery of a concept is of four types:

> Concept A. The thing only I have in mind.
> Concept B. The thing we all have in mind for a word.
> Concept C. The thing none of us may have in mind which is really the phenomenon a word stands for.
> Concept D. The unobservable phenomenon known only through its consequences which we all have in mind for a word.

An example of Concept A is when a person states an operational definition of what he alone means by a concept, "I operationalize my concept of 'nationalism' as follows . . ."

With Concept B there is an assumption that there is a shared meaning for a concept. This assumption is an empirically testable one. Social science concepts are often multidimensional complex concepts for which there is no shared meaning. Consider as examples such concepts as intelligence, nationalism, and alienation.

A typical example of Concept C is as follows: "Considerable confusion exists over the concept of political development. Although there is a general acceptance of the importance of understanding the nature of political development, there is still considerable ambiguity and imprecision in the use of the term 'political development.'" The writer is not claiming that the word refers to a phenomenon only he has in mind and therefore this is not a Concept A. Concept B is not intended since he explicitly states that there is not a shared meaning for the term and everyone does not have the same phenomenon in mind. However, the writer appears to believe that "political development" is a phenomenon, the nature of which it is important to understand. Payne (1975:36) writes:

> Another example of the same problem would be the following query:
>
> > How can we conceptualize personality integration in a way which will permit us to understand it better?
>
> Again notice how a word, "personality integration," is automatically supposed to imply a phenomenon, an "it," before any human being has seen, or defined, the phenomenon.
>
> Concept-C is the <u>verbal reification fallacy</u>: a word must have one correct meaning which is independent of the meaning human beings may have for it.

Concept D is used when we wish to postulate the existence of unobservable phenomena known only through their consequences. Such concepts are often called constructs. Payne (1975:37) gives "the force of gravity" as an example of Concept D.

> It differes from the Concept-B of "cat" in this way. If you ask me to show you what we all have in mind for "cat," I can produce the animal. If asked to produce the "force of gravity," I would say that I could not; I could produce only consequences of this force, namely objects falling down. . . . Are the two kinds of concepts, B and D, significantly different? With Concept-B, I assert that my measure identifies the phenomenon everyone has in mind for

the word "cat." With Concept-D, I assert that my measure identifies the phenomena everyone has in mind as consequences of the unobservable "force of gravity."

In the social sciences, we noted that with a Concept B there is usually not shared meaning. This is true also for a Concept D. Take as an example, alienation, a force within the individual which has certain behavioral consequences. However, there is not agreement on the behavioral consequences. Should an alienated individual join protest movements, refuse to vote, or commit suicide? Different persons may select different behavioral consequences. Of course the phenomena can be investigated. If strong dependable association is found among these consequences, then there would be reason to account for these phenomena by appealing to our unobservable concept, such as alienation.

2. Concept specification. The next step is to take the original imagery and divide it into dimensions or components. The concept that comes from the imagery consists of a complex combination of phenomena, rather than a simple and directly observable item.

An example of concept specification is Guilford's model of the structure of the intellect, which consists of 120 components. This model was proposed on the basis of more than two decades of factor-analytic research (Guilford, 1967). Guilford and his associates have identified 98 of the anticipated components (Guilford and Hoepfner, 1971). Another example is that of Kuhn and McPartland (1954) who identified the following dimensions of attitudes toward the self: favorableness, salience, consensuality, social locus, and preferences.

3. Selection of indicators. After deciding on the components or dimensions, it is necessary to find indicators for the dimensions. Take the dimension of prudence as an example. A prudent person is probably one who takes out insurance, hedges in betting, looks before he leaps,

and so on. We talk about the probability that a prudent person will perform any one of these acts compared to a less prudent person. The indicators of prudence will vary with the social setting of the individual. Students in a college organized along strict religious lines are unlikely to bet and would not have incomes that would allow them to take out insurance. Indicators of prudence relevant to the setting would be whether a student always makes a note before he lends a book, whether he never leaves his dormitory room unlocked, etc. Lazarsfeld (1958:103) states:

> The fact that each indicator has not an absolute but only a probability relation to our underlying concept requires us to consider a great many possible indicators. The case of intelligence tests furnishes an example. First, intelligence is divided into dimensions of manual intelligence, verbal intelligence, and so on. But even when there is not just one indicator by which imaginativeness can be measured, we must use many indicators to get at it.
>
> There is hardly any observation which has not at one time or another been used as an indicator of something we want to measure. We use a man's salary as one of the indicators of his ability; but we do not rely on it exclusively, or we would have to consider most businessmen more able than even top-ranking professors. We take the number of patients a doctor has cured as another indicator of ability in that setting; but we know that a good surgeon is more likely to lose a patient than is a good dermatologist. We take the number of books in a public library as an indicator of the cultural level of the community; but we know that quality of books matters as much as quantity.

A property may be a necessary, sufficient, or relevant indicator of a concept. The amount of summer rain would be a necessary but not sufficient indicator of the size of the corn crop in the Midwest of the United States. To be sufficient, other indicators would have to be taken into account, such as the amount of spring rain during the planting period. The size of the corn crop would be a sufficient but not a necessary indicator of the amount of hay fever. This is because the size

of the corn crop and the amount of ragweed, the latter being the cause of the amount of hay fever, both have common causes. The amount of ragweed would be both a necessary and a sufficient indicator of the amount of hay fever. The amount of rice eaten by a group of people is neither a necessary nor a sufficient indicator that the group of people are or are not Orientals, since people other than Orientals eat rice. Rather, the amount of rice eaten is a relevant indicator, since Orientals usually eat a lot of rice. That is, the relationship is a probability one. The most preferable type of indicator is a sufficient one, since only one indicator (and no index) is required. A necessary indicator is preferable to a relevant one, since the relationship to the concept indicated is not a probability one. A number of necessary indicators taken together may be sufficient for indicating a concept. In practice, in forming an index, it is usually possible to find only relevant indicators. In the physical sciences, in contrast to the social sciences, indices are not used because sufficient indicators of concepts have been found.

With tests, the indicators are the items, and these are commonly dichotomous. That is, if the item is correctly answered, it indicates the presence of the dimension or concept, and if not correctly answered, it indicates the absence. With dichotomous items, the test concept is in fact a classificatory concept and each item, besides being an indicator, is also a classificatory criterion. As discussed in Part I, a criterion can be related to its classificatory concept by necessity, sufficiency, or relevance. In practice, with tests, indicators (items, criteria) are related by relevance, there being no necessary or sufficient indicators of concepts. This is another way of stating "the fact that each indicator has not an absolute but only a probability relation to our

underlying concept" (Lazarsfeld, 1958:103), and consequently many indicators (items, criteria) are necessary. Indicators (items, criteria) of test concepts are chosen not only on the basis of their relevance to the test concept, but also on the basis of item analysis. The power of an indicator (item, criteria) is ascertained by correlating its value with the total test score. The size of the correlation coefficient of an indicator is greatest when subjects who obtain high total scores possess the indicator and subjects who obtain low total scores do not possess the indicator. Choosing indicators on the basis of correlation coefficients tends to result in indicators which are homogeneous. This has to be balanced against the fact that a certain degree of heterogeneity of indicators is necessary in order to capture fully the many dimensions of a test concept.

It is useful to distinguish those indicators that are part of a concept from those external to it. External criteria can furnish useful predictive criteria for the indicators which are central (most relevant) to a concept. Lazarsfeld (1958:103, 194) offers the following example:

> If we start listing indicators of the "integration" of a community, is the crime rate a part of the conception of integration, or is it an external factor which we might try to predict from our measure of integration? Here again, as with the problem of projective indices, knowing the laws which relate indicators to one another is of great importance. Even if we exclude crime rates from our image of an "integrated" city, they might be so highly correlated, as a matter of empirical generalization, that we could use them as a measure of integration in situations where we could not get data on the indicators which we "really" want to call integration. To do this, of course, we must first have "validating studies" where we correlate crime rate with other indicators of integration and establish that it is generally closely related. We should also know there are other factors besides integration influencing crime rate which might confuse our measurements if we used it alone to measure integration, so that we can check on these other factors, or add enough other indicators so as to cancel out their influence.

It is clear from Lazarsfeld's preceding statement that the choice of indicators may need to rest on rather extensive empirical knowledge of the relationship between indicators. While the initial step in the creation of our index may be some rather vague imagery, the development of the index may require quite an amount of qualitative and quantitative investigation.

4. Formation of indices. If for a particular concept, six dimensions have been identified, and ten indicators have been selected for each dimension, then in order to create an index, somehow these have to be all put together. It is somewhat like putting Humpty Dumpty together again after his having been broken into lots of pieces.

According to Zeisel (1968:84) there are two quality criteria which are pertinent to indices: accuracy and simplicity.

> By accuracy is meant the precision with which an index measures its object; and simplicity may refer either to the ease with which the necessary data can be collected or to the relative complexity of the index formula. As a rule, accuracy and simplicity compete with each other; the more an index has of the one, the less it usually has of the other.

Since in general indicators have only a probability relation or relevance to a concept, using many indicators is likely to produce a more valid and reliable index. The more indicators that are used, the more likely the different aspects of a concept are covered, thus increasing the validity. Also, the more indicators there are, the less a random fluctuation in one indicator is likely to affect the index and hence its validity and reliability. Lazarsfeld (1958:106) describes the situation as follows:

> Each indicator has only a probability relation to what we really want to know. A man might maintain his basic position, but by chance shift on an individual indicator; or he might change his basic position, but by chance remain stable on a specific indicator. But if we have

> many such indicators in an index, it is highly unlikely that a large number of them will all change in one direction, if the man we are studying has in fact not changed his basic position.
>
> To put the matter in another way, we need a lot of probings if we want to know what a man can really do or where he really stands.

The number of probings and the extent to which they cover the various aspects of a concept will affect the validity of an index. This is commonly referred to as content validity.

If there is a theory associated with a concept, this would be useful in deciding how to comgine indicators to form an index. In practice, such theories do not appear to exist, and, as already discussed, the development of an index starts from some rather vague imagery of the meaning of a conept.

One class of indices is formed by the use of multiple regression techniques. These are predictive indices, usually calculated from a linear composite of a number of variables. A common example is predicted grade point average for the first semester in college. This is taken as an index of probable success in college and is used in deciding on the entrance of students into college. Multiple regression techniques determine the nature of the index by using a sample of data in which the criterion or outcome is known. Multiple regression is purely a mathematical technique since no theory is required of how or what indicators affect the outcome. Any variable that is likely to be an indicator is tried. Even classificatory concepts such as male and female may be used by means of a technique known as dummy coding. Stepwise multiple regression techniques enable one to select a parsimonious set of indicators. That is, a linear composite of a small set of indicators which best predicts the criterion can be chosen. As might be expected

intuitively, such a set of indicators consists of indicators which have high correlation with the criterion but have low correlation with one another. The lower the correlation they have with one another, the more likely are they to comprehensively cover the range of factors that affect the outcome to be predicted. Multiple regression techniques are used quite widely to form predictive indices.

There is the problem of what meaning to attach to an index. This can be approached in two different ways. First, the composition and construction of the index can be examined to discover its meaning. Second, the index can be factor analyzed or correlated with other variables in order to reveal its meaning. This second way, in test theory, is termed construct validation and the reader is referred to the excellent paper by Cronbach and Meehl (1955) on the subject. The first way of attaching meaning to indices will be illustrated by indices of the ability of baseball hitters.

The best known index of a baseball player's hitting ability is his batting average:

$$\text{Batting average} = \frac{\text{Number of hits}}{\text{Times at bat}}$$

The batting average measures, by looking at the construction of the formula, the ability to hit the ball as often as possible.* However, frequency of hitting is not the only thing which makes for good batting. A good batter is one who not only hits as often as possible, but who hits well and when the bases are loaded. Thus, to be a good hitter, it is necessary (a) to hit often, (b) to hit well, and (c) to hit at the right time. The batting average indicates only the first of these qualities.

---

*This discussion of baseball is taken from Zeisel (1947).

An improved index would be:

$$\frac{\text{Runs batted in}}{\text{Times at bat}}$$

However, this new index contains a flaw. Batters appear usually in a fixed order at bat throughout the season, and the ones who usually hit first have, on the average, a poorer chance to bat runs in than do the batters who go to bat later, when the bases are more likely to be occupied.

A superior index is termed the slugging average. Suppose we give each hitter four points for each home run, three points for each three-base hit, two points for a two-base hit, and one point for a single, and find the total, termed "total bases."

$$\text{Slugging average} = \frac{\text{Total bases}}{\text{Times at bat}}$$

This index indicates "the number of bases" the player hits for "per time at bat." Zeisel (1968:86) comments on the meaning of slugging average:

> Note that the slugging average, because it is a more perfect measure of a batter's hitting abilty than the batting average, is in other respects an inferior index. For one thing, there are moments in a baseball game where the crucial question is not "How well will he hit?" but simply "Will he hit or will he be out?" In such situations the batting average is the superior index. The batting average has yet another advantage, namely, it is easily understood. If a player has a batting average of .333 and comes to bat, everybody knows what it means: The odds are 1 in 3 that he will get a hit. The slugging average has no such easily perceivable meaning: there is no simple way of understanding the slugging average, or of projecting it into some simple odds because it is a <u>combined</u> measure of <u>frequency</u> and <u>quality</u> of hitting, measuring the product of both. Hence, there is no way of knowing whether .340 is the result of frequent but relatively poor hitting, or of good but relatively infrequent hitting.

Zeisel's discussion illustrates the point that, since the batting average is a much simpler index than the slugging average, it is much more meaningfully interpreted. Not only that, the slugging average is a

more complex composite and it is less meaningful to attach a single number to it.

This raises the question of how valuable it is to attach a single number to a complex concept. Tests of intelligence, creativity, and scholastic aptitude are examples. Take the Wechsler Adult Intelligence Scale which is comprised of eleven subtests. The following names of the subtests indicate the host of abilities involved: information, comprehension, arithmetic, similarities, digit span, vocabulary, digit symbol, picture completion, block design, picture arrangement, and object assembly. If the slugging average is difficult to interpret because it is a combined indicator of frequency and quality of hitting, how much more difficult it is to interpret the single number of IQ based on all these abilities! Furthermore, how productive for research is it to use a single number to represent a composite of a number of such diverse abilities? An example from demography illustrates that it was not productive to attach a single value to a multidimensional variable. Medawar (1977:13) states:

> In the days when it was believed that the people of the Western world were dying out through infertility, it was thought an obligation upon demographers to devise a single value measure of a nation's reproductive prowess and future population prospectus. Kuczynski accordingly offered up his "net reproduction rate" and R. A. Fisher and A. G. Lotka the "Malthusian parameter" or "true rate of natural increase." Both had their adherents and confident predictions were based on both, but the predictions were mistaken and today no serious demographer believes that a single number valuation of reproductive vitality is feasible: reproductive vitality depends on altogether too many variables, not all of which are "scalar" in character. Among them are the proportions of married and of unmarried mothers, the prevailing fashions relating to marriage ages, family numbers, and the pattern of family building, the prevailing economic and fiscal incentives or disincentives to procreation, and the availability and social acceptability of methods of birth control. It is no wonder that the single number valuations of reproductive vitality have fallen out of

> use. Modern demographers now go about their population projections in a biologically much more realistic way, basing them essentially upon the sizes of completed familes and the analysis of "cohorts"--groups of people born or married in one specific year.

Philip (1974), in reviewing fifty years of progress in soil physics, indicates that during the early years there was a search for a single-valued physical characterization for soils. The physical properties and field behavior of soil depend upon particle size and shape, porosity, hydrogen ion concentration, material flora, water content, and hygroscopy. No single figure can embody itself in a constellation of values of all these variables in any single real instance. Candidates in the ambition to attach a single number valuation to complex variables were the hygroscopic coefficient, the wilting coefficient, the moisture equivalent, and the sticky point. The attempt to form composite variables from the values of a number of variables did not prove productive.

Indicators used in forming an index may involve any of the ways of assigning numbers. An indicator may be an index in its own right. Naming by numbers can be included since classificatory concepts, if they are named in a special way known as dummy coding, can be included in predictive indices developed by multiple regression techniques.

Ranking

Ranking is the last of the six different ways of assigning numbers to be discussed. Examination of Table 2 shows that it is the most widely-ranging way in that it can be applied in relation to all types of concepts except unordered classificatory concepts. That is, wherever there is some kind of order, ranking can be applied.

In relational to multidimensional comparative concepts, ranking can be done by the method of numerical judgment discussed in an earlier section. Indices, which also apply to multidimensional comparative concepts, can be ranked on the basis of their numerical value.

Measurements, which can be carried out only in relation to unidimensional concepts, can be ranked. Groups of entities or attributes can also be ranked according to their numerosity determined by counting. Ranking is such a straightforward process, little can be added that is not already known about it. This, then, completes the discussion of the six different ways of assigning numbers to objects and phenomena.

Number Assignment and Type of Concept

Table 2 (p. 72) summarizes Part II as well as showing that the way in which numbers can be assigned is dependent on the type of concept involved.

Naming by numbers can only be applied to classificatory concepts, whether ordered or unordered. An ordered classificatory concept involves classes which have some order to them. Ranking can be applied to any concept which has ordinal properties. Thus it can be applied to all the concepts listed in Table 2 except for unordered classificatory concepts. Comparative concepts are associated with ranking by numerical judgment in which case ranking and numerical judgment coincide. A conservative view of numerical judgment is taken by indicating that it can only be associated with comparative concepts.

Indices are applied to multidimensional comparative concepts. An index is an algebraic composite of a number of indicators. Numbers may be assigned to indicators by any one of the six ways of assigning

numbers. This includes naming numbers, as when dummy coding is used in multiple regression equations.

Measurement can only be applied to unidimensional concepts. This is because multidimensional concepts do not possess strict ordinal properties. Measurement can only be applied to unidimensional comparative, interval, and ratio concepts.

Either entities or attributes can be counted. An example of counting attributes is counting items correctly answered on a test. How a person performs on an item is an attribute of that person. Counting is applied in order to ascertain the numerosity of a group. Numerosity is an absolute concept and so counting applies to absolute concepts.

The six ways of assigning numbers are all distinctly different from one another. An example of this difference is the fact that each way is restricted in what concepts it can be applied to.

Comparison Between Physical and Social Sciences

As far as is known the social sciences, including evaluation, use all six ways of assigning numbers. This can be contrasted with the physical sciences where only measurement and counting are used. The physical sciences will occasionally use the word index to refer to some form of measurement. An example is the refractive index which is the capacity of the interface between two media to bend electromagnetic radiation. However, this index is not like the indices which has been discussed in this part, since the refractive index is not a multidimensional concept. While the refractive index is found by dividing the sine of the angle of incidence by the sine of the angle of refraction, the index could not be said to be composed of these two dimensions. As far as is known, what

TABLE 2

THE WAYS OF ASSIGNING NUMBERS TO OBJECTS AND PHENOMENA, AND THE CONCEPTS THEY ARE ASSOCIATED WITH

	Types of Concepts							
Ways of Assigning Numbers	Classificatory		Comparative					
	Un-ordered	Ordered	Multidim-ensional	Unidim-ensional	Interval	Ratio	Absolute	
Naming by Numbers	x	x						
Ranking		x	x[1]	x[1]	x	x	x	
Numerical Judgment (e.g., rating)			x	x				
Indices			x					
Measure-ment				x	x	x		
Counting: i) entities							x	
ii) attri-butes							x	

[1]Ranking of comparative concepts could be included under numerical judgment.

are termed indices in the physical sciences, are some form of measurement. Thus measurement and counting are the only two ways of assigning numbers which are found in the physical sciences.

The six ways of assigning numbers used in the social sciences are frequently all termed "measurement" and no clear terminological difference is made between them. It becomes clear that what is termed quantification in the social sciences is different from that in the physical sciences. However, many social scientists believe that the key to success in the physical sciences is quantification and this is also the key to the success of the social sciences. If the physical sciences use only measurement and counting, then the additional four forms of assigning numbers used by the social sciences may not be the key to success.

In Part I classificatory concepts were not regarded as quantitative concepts. Thus from Table 2 it can be seen that naming by numbers is not a form of quantification and this is also true when ranking is used with classificatory concepts. Numerical judgment and indices can be regarded as forms of quantification. While both of these are useful, they do have their limitations. Reliability and validity is a problem with numerical judgment. A serious problem with indices is knowing how to interpret them. Numerical judgments and indices are the two forms of quantification most commonly used in the social sciences. Measurement, in the physical science sense, is not common in the social sciences.

PART III: THE HISTORY OF THE QUANTITATIVE AND QUALITATIVE IN THE PHYSICAL SCIENCES AND THE IMPLICATIONS FOR EVALUATION

At the University of Chicago, the facade of the Social Science Research Building bears the dictum,

> If you cannot measure, your knowledge
> is meager and unsatisfactory.

While one might expect that it was written by a social scientist, it was not. The authorship belongs to one of the giants of the physical sciences, Lord Kelvin. The placing of the dictum does illustrate that, to social scientists in general, physical science is so often seen as the paradigm of sound knowledge, and quantitative techniques seem to provide an essential clue to its success. Unfortunately social scientists often see the physical sciences as they are now, with their highly developed conceptual networks. Quantitative techniques are extremely powerful tools. But it was not always so. In the beginning and through the long centuries, a base of qualitative knowledge had to be built up before conceptual frameworks on which measurement rests could be developed. Kuhn (1961, p. 31) feels that it has only been for the last century and a half that quantitative methods have been central to the physical sciences. The following discussion of quantification in modern physical science is based on the work of Kuhn (1961). His central thesis is that large amounts of qualitative work have usually been prerequisite to fruitful quantification in the physical sciences. Evaluation draws on the social sciences. Since the social sciences are in the early stages of development, this suggests that quantification in evaluation may not be as fruitful as qualitative methodology.

The View of Quantification from Science Texts

For most of us, our view of physical science and quantification comes from science textbooks. The view of theory and quantification presented in the science textbooks can be summarized as follows.

Theory and Laws → Logical and Mathematical Manipulation → Results

Theory	Experiment
1.361	1.368
1.919	1.925
2.432	2.434
2.849	2.850

The left of the diagram represents a series of theoretical and lawlike statements which together constitute the theory of the science being described. The center of the diagram represents the logical and mathematical equipment employed in manipulating the theory. The logical and mathematical manipulations are carried out on the conditions of the situation to which the theory is applied. The result is a set of numerical predictions shown on the left-hand side of the table. The right-hand side of the table shows the numerical results of actual measurements, placed there so that they may be compared with the predictions derived from theory. Many textbooks in physics, chemistry, astronomy, etc., have data of this kind, though they are not always presented in tabular form. Sometimes they are presented in graphical form.

Evaluation draws on the social sciences. Social science textbooks contain numerical tables which are of a different type. The numerical tables are used to illustrate the relationship between a dependent variable and one or more independent variables. To take a simple example, there might be two columns of figures showing math achievement, one column for Grade 12 boys and one for Grade 12 girls. The function of the table is to compare the achievement of the boys and girls and to show

that the achievement of the boys is significantly greater (statistically) than for the girls. That is, math achievement is dependent on sex of the student. This is quite different from the tables in the physical sciences where a theory predicts numerical results and experimental results are compared to them.

What is the function of such tables of numbers in the physical sciences? There are at least three answers.

1. The most obvious answer is that the numbers in the table function as a test of the theory. If the corresponding numbers in the two columns agree, then the theory is accepted. If they do not agree, then the theory is rejected. Thus, the function of quantification in the physical sciences would appear to be one of confirmation. This might be so in the practice of science but it is unlikely to be so in a textbook formulation. No textbook ever included a table that either intended c managed to challenge the adequacy of the theory described by the text. Readers of current science texts accept the theories there expounded on the authority of the author and the scientific community, not because of any tables that these texts contain. If the tables are read at all, as they often are, they are read for another reason.

In contrast, in the social science textbooks, the tables are there to illustrate confirmation of some hypothesis, a hypothesis being more tentative than a theory. In order to back up the hypothesis, in addition to the table, it is common to quote a number of authors who have also found evidence for the hypothesis.

2. A common belief is that the function of quantification in the physical sciences is for exploration. Like the function of confirmation, Kuhn (1961, pp. 34, 35) is skeptical that exploration is a function of the numerical tables of the textbooks. He states:

> Numerical data like those collected in the right-hand column of our table can, it is often supposed, be useful in suggesting new scientific theories or laws. Some people seem to take for granted that numerical data are more likely to be productive of new generalizations than any other sort. It is that special productivity, rather than measurement's function in confirmation, that probably accounts for Kelvin's dictum being inscribed on the facade at the University of Chicago. . . . We are, I suspect, here confronted with a vestige of an admittedly outworn belief that laws and theories can be arrived at by some process like "running the machine backwards." Given the numerical data in the "Experiment" column of the table, logico-mathematical manipulation (aided, all would now insist, by "intuition") can proceed to the statement of the laws that underlie the numbers. If any process even remotely like this is involved in discovery--if, that is, laws and theories are forged directly from data by the mind--then the superiority of numerical to qualitative data is immediately apparent. The results of measurement are neutral and precise; they cannot mislead. Even more important, numbers are subject to mathematical manipulation; more than any other form of data, they can be assimilated to the semimechanical textbook schema.

An example of a data analysis technique in the social sciences that attempts to move from numerical data to theory is factor analysis. By analyzing a correlation matrix it is hoped to discover the theoretical factors that are responsible for the data. However, factor analysis can also be used in a confirmatory mode, i.e., to confirm some postulated factor structure.

Tukey (1976) has recently written a book systematizing statistical techniques for exploratory data analysis in the social sciences. The aim of the techniques appears to be to discover the regularity or generalization that lies behind a set of data. Tukey is not rejecting confirmatory statistical techniques, for he states that the exploratory and the confirmatory can--and should--proceed side by side.

3. One cannot expect that theoretical results will agree with experimental results. There are a number of reasons for this. One reason is that there is always limitations to the accuracy of the measuring instruments employed. In addition computation from theory can

usually be carried out to any desired number of decimal places, and this makes agreement between theoretical and experimental results impossible. Another reason is that the theory may make some simplifying assumptions about the world. Almost always the application of a physical theory involves some approximation (in fact, the plane is not "frictionless," the vacuum is not "perfect," the atoms are not "unaffected" by collisions), and the theory is not therefore expected to yield quite precise results. Thus, what scientists seek in numerical tables is not usually "agreement" at all, but what they often call "reasonable agreement." Kuhn (1961, p. 36) believes that, when they appear in a text, tables of numbers drawn from theory and experiments define for the reader what is reasonable agreement. Kuhn (1961, p. 36) states:

> That, I think, is why the tables are there: they define "reasonable agreement." By studying them, the reader learns what can be expected of the theory. An acquaintance with the tables is part of an acquaintance with the theory itself. Without the tables, the theory would be essentially incomplete. With respect to measurement, it would be not so much untested as untestable. Which brings us close to the conclusion that, once it has been embodied in a text--which for present purposes means, once it has been adopted by the profession--no theory is recognized to be testable by any quantitative tests that it has not already passed.

This is different for the social sciences. The inclusion of experimental data in a textbook to support some hypothesis does not mean that the hypothesis has generally been adopted by the profession. The experimental data is there to illustrate confirmation of the hypothesis by a researcher or researchers. The theories of the social sciences are not of the nature that they supply numerical predictions. So there are no tables to illustrate what is "reasonable agreement" between theoretical and experimental results.

Motives for Normal Measurement

In his most influential work, The Structure of Scientific Revolutions, Kuhn (1962) has as his basic problem the nature of scientific change. In summary, his thesis is that "scientific revolutions are . . . those non-cumulative developmental episodes in which an older paradigm is replaced in whole or in part by an incompatible new one" (p. 91), where paradigms are defined to be "accepted examples of actual scientific practice--examples which include law, theory, application, and instrumentation together--[which] provide models from which spring particular coherent traditions of scientific research" (p. 10). The work of Einstein is a dramatic example of revolutionary (or extraordinary) science. The new Einsteinian paradigm replaced the older Newtonian one. Often during a revolutionary period there is competition between paradigms or theories until one is finally accepted.

If scientific change is fundamentally revolutionary, there must be nonrevolutionary periods as well and Kuhn calls this nonrevolutionary science, normal science. During a period of normal science there is fundamental agreement within the scientific community. What is characteristic of normal science is that it is carried out by a scientific community which shares "firm answers to questions like the following: What are the fundamental entities of which the universe is composed? How do these interact with each other and with the senses? What questions may be legitimately asked about such entities and what techniques employed in seeking solutions?" (Kuhn, 1962, pp. 4-5). In addition the members of such a scientific community also share common values. Revolutionary new theories are relatively rare and most of the

science engaged in is normal science. Kuhn (1961, p. 38) explains how normal science follows on from revolutionary science.

> The new order provided by a revolutionary new theory in the natural sciences is always overwhelmingly a potential order. Much work and skill, together with occasional genius, are required to make it actual. And actual it must be made, for only through the process of actualization can occasions for new theoretical reformulations be discovered. The bulk of scientific practice is thus a complex and consuming mopping-up operation that consolidates the ground made available by the most recent theoretical breakthrough and that provides essential preparation for the breakthrough to follow. In such mopping-up operations, measurement has its overwhelmingly most common scientific function.

Quantification during a period of normal science involves comparing the numerical predictions of theory with the actual theoretical measurements. The problem that engaged much of the best eighteenth-century scientific thought was that of deriving numerical predictions from Newton's three Laws of Motion and from his principle of universal gravitation and comparing them with experimental measurements. Again the problem arose of what would be reasonable agreement, taking into account one could not obtain perfect experimental conditions. As an example, consider the application of Newtonian mechanics to the pendulum. The suspensions of laboratory pendula are neither weightless nor perfectly elastic; air resistance damps the motion of the bob; besides, the bob itself is of finite size, and there is the question of which point of the bob should be used in computing the pendumlum's length. If these three aspects of the experimental situation are neglected, only the roughest sort of quantitative agreement between theory and observation can be expected. But determining how to reduce them (only the last is fully eliminable) and what allowance to make for the residue are themselves of the utmost difficulty. Since Newton's day much brilliant research has been devoted to their challenge. Kuhn (1961, p. 41) states:

> This is the sort of work that most physical scientists do most of the time insofar as their work is quantitative. Its objective is, on the one hand, to improve the measure of "reasonable agreement" applicable to them. . . . If measurement ever leds to discovery or to confirmation, it does not do so in the most usual of all its applications."

According to Kuhn theories are not tested during a period of normal science, only during revolutionary science. During normal science when a scientist finds agreement between theoretical and experimental results this is not seen as a confirmation of the theory. The scientist's success lies only in the explicit demonstration of a previously implicit agreement between theory and the world. No novelty in nature has been revealed. Failure to obtain agreement is seen as counting against the scientist; his talents were not adequate to the task of obtaining reasonable agreement.

Evaluation is different from the science Kuhn is talking about. Evaluation is a practical activity whereas science aims to develop theories of phenomena that aid understanding. Strike (1979, p. 13) claims that there are practical theories corresponding to the explanatory theories of science.

> . . . research in practical areas is not atheoretical. Practical and theoretical are not properly opposed terms. Rather we need to distinguish between explanatory theories--those concerned with understanding, and practical theories--those concerned with action. The latter are much like explanatory theories in that they can guide practical inquiry as explanatory theories guide their sort of research. In both cases they need to be mapped onto experience. Moreover, they exhibit similar patterns of conceptual change.

Kuhn's theory of scientific revolutions is an example of a theory of conceptual change. There is a possibility that measurement may function differently with an explanatory theory compared to a practical theory.

Is evaluation in 1979 in a preparadigmatic stage, a normal stage or a revolutionary stage? An argument can be made for each one of these

three positions. The following is a brief summary of each of the three positions.

The preparadigmatic stage occurs in the initial development of a science before it acquires its first universally received paradigm (Kuhn, 1970, p. 13). In this stage there is a number of competing schools or paradigms. Worthen and Sanders (1973), in giving a history of educational program evaluation, seem to indicate that educational evaluation did not begin in earnest till the advent of the Elementary and Secondary Education Act of 1965 (ESEA). It was required that each project under Titles I and III of the ESEA be evaluated. Educators were unprepared to implement the new mandate effectively. Worthen and Sanders (1973, p. 6) state:

> Few scholars had turned their attention to the development of generalizable evaluation plans which could be adopted or adapted by local evaluators. Theoretical work in evaluation was almost nonexistent. However, scholars--like nature--abhor a vacuum, and it was not long before several evaluation theoreticians began to develop and test their notions about how one should conduct educational evaluations. These efforts resulted in several new evaluation models, strategies, and plans which could be put into use by educationists.

Typically in the preparadigmatic stage there are a number of paradigms or theories. Different writers list different numbers of practical theories in evaluation. Stake (1976, p. 28) lists nine. Worthen and Sanders (1973, pp. 210-215) indicate eight theories while House (1978, p. 12) lists a different set of eight. It would appear that evaluation has not stabilized by the acceptance of a single paradigm.

An argument could be put forward that evaluation is in a stage of normal science. Kuhn (1970, p. 11) states:

> The study of paradigms . . . is what mainly prepares the student for membership in the particular scientific community with which he will later practice. Because he there joins men who learned the bases of their field from

> the same concrete models, his subsequent practice will seldom evoke overt disagreement over fundamentals. Men whose research is based on shared paradigms are committed to the same rules and standards for scientific practice. That commitment and the apparent consensus it produces are prerequisites for normal science, i.e., for the genesis and continuation of a particular research tradition.

One place to look for evidence of a possible consensus is at graduate training in evaluation. In 1976, 81 graduate schools of education that offer degree programs above the masters level replied to a questionnaire requesting information on their evaluation training programs (Gephart and Potter, 1976). This project was carried out for the Evaluation Network, a professional society of evaluators. From school to school the course offerings were consistently centered around statistical analysis, research design and testing. There is clearly a dominant methodological paradigm which is quantitative. The practical theory behind this methodology is that a social action program is a form of treatment with inputs and outputs. A research design is set up. The inputs and outputs are measured often with the help of tests. Some form of comparison is then made via statistical analysis of the measurements. It is clear that the purpose of measurement in evaluation in a normal stage is quite different from measurement in a period of normal science. The dominant practical theory of evaluation, while it is quantitative, does not make numerical predictions. It is not a matter of comparing experimental results with theoretical predictions as is the case with normal science.

It could be argued that evaluation is moving into a revolutionary stage. Within the last few years there has been advocacy and use of qualitative methods for evaluation (e.g., Parlett and Hamilton, 1972; Stake and Easley, 1978; Guba, 1978). The current quantitative paradigm has been criticized as being inadequate. However, the number of

adherents to the quantitative paradigm is still large and it is doubtful if it will be replaced by a qualitative paradigm in the near future.

The Effects of Normal Measurement.

It is usually assumed that the physical scientist works by making theories conform to the measurements obtained. In fact the reverse is the case. In a period of normal science the physical scientist works to make his measurements conform to his theory. In other words, a theory is not derived from a set of quantitative measurements. Unless the theory is known beforehand, it is not obvious from a set of numbers what the theory is. Kuhn (1961, pp. 45, 46) states:

> Numbers gathered without some knowledge of the regularity to be expected almost never speak for themselves. Almost certainly they remain just numbers. This does not mean that no one has ever discovered a quantitative regularity merely by measuring. . . . But, partly just because they are so exceptional and partly because they never occur until the scientist measuring knows everything but the particular form of the quantitative result he will obtain, these exceptions show just how improbably quantitative discovery by quantitative measurement is. . . . [a large] amount of theory is needed before the results of measurement can be expected to make sense. But, and this is perhaps the main point, when that much theory is available, the law is very likely to have been guessed without measurement.

If theory precedes quantitative measurements, then how is the theory formed? A theory begins in a qualitative form and only later does its quantitative implications become evident. The development of a theory depends on qualitative data. It is qualitative experimentation that dominates the earlier developmental stages of a physical science and continues to play a role later on. Significant quantitative comparison of a theory with nature comes at the late stage of the development of a science. The early development of a theory is likely to require insight and conceptualization.

Evaluators draw on the theories and methodologies of the social sciences. The social sciences, while they are in the early stages of their development, have used quantitative experimentation extensively. However, this has not led to quantitative theories. If the social sciences in any way parallel the physical sciences, quantitative theories are only likely to arise out of a firm base of qualitative knowledge. This suggests that at this early stage of their development, qualitative rather than quantitative experimentation is likely to be more productive in the social sciences. Correspondingly, this also suggests that qualitative methods are likely to be more productive in evaluation. This is especially so when an evaluator enters an unfamiliar situation which an evaluator is attempting to understand.

Revolutionary Science

During a period of revolutionary science one theory supersedes another. According to Kuhn (1961), measurement plays an important part during this revolutionary stage. Though there are no apparent implications for evaluation, measurement during a revolution stage will be discussed for the sake of completeness.

In a revolutionary stage, science is in a state of crisis. There is an unacceptable anomaly between established theory and experiment. A state of dissatisfaction with established theory arises. According to Kuhn (1961, p. 52), no crisis is so hard to suppress as one that derives from a quantitative anomaly that has resisted all the usual efforts at reconciliation. That is, there is an unexplained discrepancy between a measurement and the value predicted by theory. Kuhn (1961, pp. 53, 54) gives a number of examples

> "to illustrate how difficult it is to explain away established quantitative anomalies, and to show how much more effective these are than qualitative anomalies in establishing inevitable scientific crises. But the examples also show something more. They indicate that measurement can be an immensely powerful weapon in the battle between two theories, and that, I think, is its second particularly significant function. Furthermore, it is for this function--aid in the choice between theories--and for it alone, that we must reserve the word "confirmation."

Measurement in the Development of Physical Science

The history of the physical sciences indicates that much qualitative work, both empirical and theoretical, is necessary before quantification proves to be productive. A good example is the field of optics. The seventeenth century's Scientific Revolution's reformulation of optical theory depended on Newton's prism experiments. These experiments depended on much prior qualitative work. Newton's innovation was the quantitative analysis of a well-known qualitative effect. Another example is that of magnetism. The only significant seventeenth-century measurements, those of declination and dip, were made with one or another modified version of the traditional compass, and these measurements did little to improve the understanding of magnetic phenomena. For a more fundamental quantification, magnetism awaited the work of Coulomb, Gauss, Poisson, and others in the late eighteenth and early nineteenth centuries. Before that work could be done, a better qualitative understanding of attraction, repulsion, conduction, and other such phenomena was needed. The instruments which produced a lasting quantification had then to be designed with these initially qualitative conceptions in mind. Kuhn (1961, pp. 55, 60) states:

> . . . much qualitative research, both empirical and theoretical, is normally prerequisite to fruitful quantification of a given research field. In the absence of such prior work, the methodological directive, "Go ye

> forth and measure," may well prove only an invitation to waste time. . . . I venture the following paradox: The full and intimate quantification of any science is a consummation devoutly to be wished. Nevertheless, it is not a consummation that can effectively be sought by measuring. As in individual development, so in the scientific group, maturity comes most surely to those who know how to wait.

Since the social sciences are in an early stage of development, it could be expected that qualitative methods rather than quantification would predominate. In reality, the reverse is true. Thus it can be wondered to what extent quantification in the social sciences is a waste of time.

One point to notice is that quantification in the social and physical sciences are essentially different. Quantification in the physical sciences is synonymous with measurement. In the social sciences, besides measurement, other forms of quantification are used such as ranking, rating, indices and counting. In the social sciences, the word "measurement" is used to cover these other forms of quantification. As an example, tests do not measure in the physical science sense but result in indices. Questionnaires commonly require respondents to rate which is different from the process of measurement in the physical science. In fact very little of the physical science type of measurement occurs in the social sciences. Thus the social sciences may not be wasting time as Kuhn's thesis would suggest. On the other hand, the use of ranking, rating, indices and counting is not aimed at qualitative understanding which, according to Kuhn, precedes fruitful measurement (in the physical science sense).

References

Anastasi, A. Psychological Testing. Fourth Edition. New York: Macmillan, 1976.

Bonjean, C. M., Hill, R. J., and McLemore, S. D. Sociological Measurement. San Francisco: Chandler, 1967.

Boorer, M. Mammals of the world. New York: Gorsset & Dunlap, 1979.

Campbell, N. R. Symposium: Measurement and its importance for philosophy. Proceedings of the Aristotelian Society, Supplementary Volume 17, 121-142. London: Harrison, 1938.

Campbell, N. R. Physics, the elements. London: Cambridge University Press, 1920.

Carnap, R. Logical foundations of probability. 2nd Edition. Chicago: University of Chicago Press, 1962.

Carrington, R. The mammals. New York: Time Incorporated, 1963.

Cockrum, E. L. Introduction to mammalogy. New York: The Ronald Press Company, 1962.

Christy, G. A., and Clendenin, J. C. Introduction to investments. New York: McGraw-Hill, Sixth Edition, 1974.

Churchman, C. W. Why measure? In C. W. Churchman and P. Ratoosh (Eds.) Measurement: Definitions and theories. New York: Wiley, 1959.

Cohen, Mr. R., and Nagel, E. An introduction to logic and scientific method. New York: Harcourt, Brace & World, 1934.

Coombs, C. H., Dawes, R. M., and Tversky, A. Mathematical psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Cronbach, L. J. On the design of educational measures. In D. De Gruijter and L. van der Kamp (Eds.) Advances in psychological and educational measurement. London: Wiley, 1976.

Cronbach, L. J., and Gleser, G. C. Psychological tests and personnel decisions. Urbana, Ill.: University of Illinois Press, 1965.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Cronbach, L. J., and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

DeBlase, A. F., and Martin, R. E. A manual of mammalogy. Dubuque, Iowa: Wm. C. Brown, 1974.

Ellis, B. Basic concepts of measurement. London: Cambridge University Press, 1966.

Gephart, W. J. and Potter, W. J. Evauation training catalog. Bloomington, Indiana: Phi Delta Kappa, 1976.

Gilmour, J. S. Taxonomy and philosophy. In J. Huxley (Ed.) The new systematics. London: Oxford University Press, 1960.

Green, T. F. The activities of teaching. New York: McGraw-Hill, 1979.

Guba, E. G. Toward a methodology of naturalistic inquiry in educational evaluation. Los Angeles: Center for the Study of Evaluation, University of Californiua, 1978.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill 1967.

Guilford, J. P., and Hoepfner, R. The analysis of intelligence. New York: McGraw-Hill, 1971.

Hagood, M. J., and Price, D. O. Statistics for sociologists. New York: Henry Holt, 1952.

Hanson, N. R. Perception and discovery. San Francisco: Freeman, Cooper & Co., 1969.

Hempel, C. G. Fundaments of concept formation in empirical science. International Encyclopedia of Unified Science, Vol. 2, No. 7. Chicago: University of Chicago Press, 1952.

Hoffmeister, D. F. Mammals. New York: Golden Press, 1963.

House, E. R. Assumptions underlying evaluation models. Educational Researcher, 1978, 7, No. 3, 4-12.

Jones, L. V. The nature of measurement. In R. L. Thorndike (Ed.) Educational measurement. Washington, D.C.: American Council on Education, 1971.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. Foundations of measurement, Volume I. New York: Academic Press, 1971.

Kuhn, M. H., and McPartland, T. S. An empirical investigation of self attitudes. American Sociological Review, 1954, 19, 68-76.

Kuhn, T. S. The function of measurement in modern physical science. In H. Woolf (Ed.) Quantification. Indianapolis: Bobbs-Merrill, 1961.

Kuhn, T. S. The structure of scientific revolutions. Chicago: University of Chicago Press, 1st edit. 1962, 2nd edit. 1970.

Lazarsfeld, P. F. Evidence and inference in social research. Daedalus, 1958, 87, 99-130.

Lazarsfeld, P. F., and Rosenberg, M. The language of social research. Glencoe, Ill.: The Free Press, 1955.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addision-Wesley, 1968.

Lumsden, J. Test theory. In M. R. Rosenzweig and L. W. Porter (Eds.) Annual Review of Psychology, Vol. 27. Palo Alto, Cal.: Annual Reviews, 1976.

Mayr, E. Systematics and the origin of species. New York: Columbia University Press, 1942.

Medawar, P. B. Unnatural science. The New York Review, Feb. 3, 1977, 13-18.

Morris, D. The mammals. New York: Harper & Rowe, 1965.

Nagel, N. Measurement. In A. Danto and S. Morgenbesser (Eds.) Philosophy of science. New York: The World Publishing Company, 1960.

Nelson, T. M., and Bartley, S. H. Numerosity, number, arithmetization, measurement and psychology. Philosophy of Science, 1961, 28, 178-203.

Parlett, M. and Hamilton, D. Evaluation as illumination: a new approach to the study of innovatory programs. Edinburgh: University of Edinburg, 1972.

Payne, J. L. Principles of social science measurement. College Station, Texas: Lytton, 1975.

Philip, J. R. Fifty years progress in soil physics. Geoderma, 1974, 12, 265-280.

Polanyi, M. Personal knowledge. New York: Harper & Rowe, 1958.

Russell, B. Principles of mathematics. Second Edition. New York: Norton, 1938.

Siegel, S. Nonparametric statistics. New York: McGraw-Hill, 1956.

Smith, B. O. Logical aspects of educational measurement. New York: Columbia University Press, 1938.

Stake, R. E. Evaluating educational programmes. Paris: OECD, 1976.

Stake, R. E. and Easley, J. A. Case studies in science education. Urbana-Champaign: University of Illinois, 1978.

Stevens, S. S. Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.) Handbook of experimental psychology. New York: Wiley, 1951.

Stevens, S. S. Measurement, psychophysics, and utility. In C. W. Churchman and P. Ratoosh (Eds.) Measurement: Definitions and theories. New York: Wiley, 1959.

Stibbe, E. P. Physical anthropology. London: Edward Arnold, 1930.

Strike, K. A. An epistemology of practical research. Educational Researcher, 1979, 8, No. 1, 10-16.

Suppes, P.., and Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.) Handbook of mathematical psychology, Volume I. New York: Wiley, 1963.

Torgerson, W. S. Theory and method of scaling. New York: Wiley, 1958.

Toulmin, S., and Goodfield, J. The fabric of the heavens. New York: Harper & Rowe, 1961.

Tukey, J. W. Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley, 1977.

Worthen, B. R. and Sanders, J. R. Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Zeisel, H. Say it with figures. First Edition. New York: Harper & Brothers, 1947.

Zeisel, H. Say it with figures. Fifth Edition. New York: Harper & Rowe, 1968.